

A Comprehensive Introduction to Differential Geometry
3rd edition (Publish or Perish, 1999)

Appendix: *Calculus on Manifolds* (Addison-Wesley, 1968)

by

Michael Spivak

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME ONE
Third Edition



MICHAEL SPIVAK

PUBLISH OR PERISH, INC.



Houston, Texas 1999

ACKNOWLEDGEMENTS

I am greatly indebted to

Richard S. Palais

without his encouragement
these volumes would have remained
a short set of mimeographed notes

and

Donald E. Knuth

without his T_EX program they would
never have become typeset books

PREFACE

The preface to the first edition, reprinted on the succeeding pages, excused this book's deficiencies on grounds that can hardly be justified now that these "notes" truly have become a book.

At one time I had optimistically planned to completely revise all this material for the momentous occasion, but I soon realized the futility of such an undertaking. As I examined these five volumes, written so many years ago, I could scarcely believe that I had once had the energy to learn so much material, or even recall how I had unearthed some of it.

So I have contented myself with the correction of errors brought to my attention by diligent readers, together with a few expository ameliorations; among these is the inclusion of a translation of Gauss' paper in Volume 2.

Aside from that, this third and final edition differs from the previous ones only in being typeset, and with figures redrawn. I have merely endeavored to typeset these books in a manner befitting a subject of such importance and beauty.

As a final note, it should be pointed out that since the first volumes of this series made their appearance in 1970, references in the text to "recent" results should be placed in context.

Preface to the First Edition

HOW THESE NOTES CAME TO BE

and how they did not come to be a book

For many years I have wanted to write the Great American Differential Geometry book. Today a dilemma confronts any one intent on penetrating the mysteries of differential geometry. On the one hand, one can consult numerous classical treatments of the subject in an attempt to form some idea how the concepts within it developed. Unfortunately, a modern mathematical education tends to make classical mathematical works inaccessible, particularly those in differential geometry. On the other hand, one can now find texts as modern in spirit, and as clean in exposition, as Bourbaki's Algebra. But a thorough study of these books usually leaves one unprepared to consult classical works, and entirely ignorant of the relationship between elegant modern constructions and their classical counterparts. Most students eventually find that this ignorance of the roots of the subject has its price -- no one denies that modern definitions are clear, elegant, and precise; it's just that it's impossible to comprehend how any one ever thought of them. And even after one does master a modern treatment of differential geometry, other modern treatments often appear simply to be about totally different subjects.

Of course, these remarks merely mean that no matter how well some of the present day texts achieve their objective, I nevertheless feel that an introduction to differential geometry ought to have quite different aims. There are two main premises on which these notes are based. The first premise is that it is absurdly inefficient to eschew the modern language of manifolds, bundles, forms, etc., which was developed precisely in order to rigorize the concepts of classical differential geometry. Rephrasing everything in more elementary terms involves incredible

contortions which are not only unnecessary, but misleading. The work of Gauss, for example, which uses infinitesimals throughout, is most naturally rephrased in terms of differentials, even if it is possible to rewrite it in terms of derivatives. For this reason, the entire first volume of these notes is devoted to the theory of differentiable manifolds, the basic language of modern differential geometry. This language is compared whenever possible with the classical language, so that classical works can then be read.

The second premise for these notes is that in order for an introduction to differential geometry to expose the geometric aspect of the subject, an historical approach is necessary; there is no point in introducing the curvature tensor without explaining how it was invented and what it has to do with curvature. I personally felt that I could never acquire a satisfactory understanding of differentiable geometry until I read the original works. The second volume of these notes gives a detailed exposition of the fundamental papers of Gauss and Riemann. Gauss' work is now available in English (General Investigations of Curved Surfaces; Raven Press). There are also two English translations of Riemann's work, but I have provided a (very free) translation in the second volume.

Of course, I do not think that one should follow all the intricacies of the historical process, with its inevitable duplications and false leads. What is intended, rather, is a presentation of the subject along the lines which its development might have followed; as Bernard Morin said to me, there is no reason, in mathematics any more than in biology, why ontogeny must recapitulate phylogeny. When modern terminology finally is introduced, it should be as an outgrowth of this (mythical) historical development. And all the major approaches have to be presented, for they were all related to each other, and all still play an important role.

At this point I am reminded of a paper described in Littlewood's Mathematician's Miscellany. The paper began "The aim of this paper is to prove ..." and it transpired only much later that this aim was not achieved (the author hadn't claimed that it was). What I have outlined above is the content of a book the realization of whose basic plan and the incorporation of whose details would perhaps be impossible; what I have written is a second or third draft of a preliminary version of this book. I have had to restrict myself to what I could write and learn about within the present academic year, and all revisions and corrections have had to be made within this same period of time. Although I may some day be able to devote to its completion the time which such an undertaking deserves, at present I have no plans for this. Consequently, I would like to make these notes available now, despite their deficiencies, and with all the compromises I learned to make in the early hours of the morning.

These notes were written while I was teaching a year course in differential geometry at Brandeis University, during the academic year 1969-70. The course was taken by six juniors and seniors, and audited by a few graduate students. Most of them were familiar with the material in Calculus on Manifolds, which is essentially regarded as a prerequisite. More precisely, the complete prerequisites are advanced calculus using linear algebra and a basic knowledge of metric spaces. An acquaintance with topological spaces is even better, since it allows one to avoid the technical troubles which are sometimes relegated to the Problems, but I tried hard to make everything work without it.

The material in the present volume was covered in the first term, except for Chapter 10, which occupied the first couple of weeks of the second term, and Chapter 11, which was not covered in class at all. We found it necessary to take rest cures of nearly a week after completing Chapters 2, 3, and 7. The same material could easily be expanded to a full year course

in manifold theory with a pace that few would describe as excessively leisurely. I am grateful to the class for keeping up with my accelerated pace, for otherwise the second half of these notes would not have been written. I am also extremely grateful to Richard Palais, whose expert knowledge saved me innumerable hours of labor.

*Michael Spivak
Brandeis University
March, 1970*

TABLE OF CONTENTS

Although the chapters are not divided into sections,
the listing for each chapter gives some indication
which topics are treated, and on what pages.

CHAPTER 1. MANIFOLDS

Elementary properties of manifolds	1
Examples of manifolds	6
Problems	20

CHAPTER 2. DIFFERENTIAL STRUCTURES

C^∞ structures	27
C^∞ functions	31
Partial derivatives	35
Critical points	40
Immersion theorems	42
Partitions of unity	50
Problems	53

CHAPTER 3. THE TANGENT BUNDLE

The tangent space of \mathbb{R}^n	63
The tangent space of an imbedded manifold	67
Vector bundles	71
The tangent bundle of a manifold	75
Equivalence classes of curves, and derivations	77
Vector fields	82
Orientation	84
Addendum. Equivalence of Tangent Bundles	89
Problems	95

CHAPTER 4. TENSORS

The dual bundle	107
The differential of a function	109
Classical <i>versus</i> modern terminology	111
Multilinear functions	115
Covariant and contravariant tensors	117
Mixed tensors, and contraction	121
Problems	127

CHAPTER 5. VECTOR FIELDS AND DIFFERENTIAL EQUATIONS

Integral curves	135
Existence and uniqueness theorems	139
The local flow	143
One-parameter groups of diffeomorphisms	148
Lie derivatives	150
Brackets	153
Addendum 1. Differential Equations	164
Addendum 2. Parameter Curves in Two Dimensions	167
Problems	169

CHAPTER 6. INTEGRAL MANIFOLDS

Prologue; classical integrability theorems	179
Local Theory; Frobenius integrability theorem	190
Global Theory	194
Problems	198

CHAPTER 7. DIFFERENTIAL FORMS

Alternating functions	201
The wedge product	203
Forms	207
Differential of a form	210
Frobenius integrability theorem (second version)	215
Closed and exact forms	218
The Poincaré Lemma	225
Problems	227

CHAPTER 8. INTEGRATION

Classical line and surface integrals	239
Integrals over singular k -cubes	246
The boundary of a chain	248
Stokes' Theorem	253
Integrals over manifolds	256
Volume elements	258
Stokes' Theorem	261
de Rham cohomology	263
Problems	283

CHAPTER 9. RIEMANNIAN METRICS

Inner products	301
Riemannian metrics	308
Length of curves	312
The calculus of variations	316
The First Variation Formula and geodesics	323
The exponential map	334
Geodesic completeness	341
Addendum. Tubular Neighborhoods	344
Problems	348

CHAPTER 10. LIE GROUPS

Lie groups	371
Left invariant vector fields	374
Lie algebras	376
Subgroups and subalgebras	379
Homomorphisms	380
One-parameter subgroups	382
The exponential map	384
Closed subgroups	391
Left invariant forms	394
Bi-invariant metrics	400
The equations of structure	402
Problems	406

CHAPTER 11. EXCURSION IN THE REALM OF ALGEBRAIC TOPOLOGY

Complexes and exact sequences	419
The Mayer-Vietoris sequence	424
Triangulations	426
The Euler characteristic	428
Mayer-Vietoris sequence for compact supports	430
The exact sequence of a pair	432
Poincaré Duality	439
The Thom class	442
Index of a vector field	446
Poincaré-Hopf Theorem	450
Problems	453

APPENDIX A

To Chapter 1	459
Problems	467
To Chapter 2	471
Problems	472
To Chapter 6	473
To Chapters 7, 9, 10	474
Problem	475

NOTATION INDEX	477
--------------------------	-----

INDEX	481
-----------------	-----

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME ONE

CHAPTER 1

MANIFOLDS

The nicest example of a metric space is Euclidean n -space \mathbb{R}^n , consisting of all n -tuples $x = (x^1, \dots, x^n)$ with each $x^i \in \mathbb{R}$, where \mathbb{R} is the set of real numbers. Whenever we speak of \mathbb{R}^n as a metric space, we shall assume that it has the “usual metric”

$$d(x, y) = \sqrt{\sum_{i=1}^n (y^i - x^i)^2},$$

unless another metric is explicitly suggested. For $n = 0$ we will interpret \mathbb{R}^0 as the single point $0 \in \mathbb{R}$.

A manifold is supposed to be “locally” like one of these exemplary metric spaces \mathbb{R}^n . To be precise, a manifold is a metric space M with the following property:

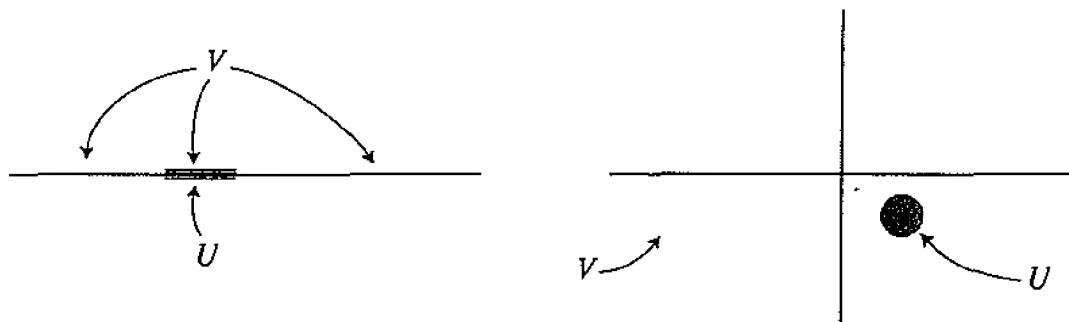
If $x \in M$, then there is some neighborhood U of x and some integer $n \geq 0$ such that U is homeomorphic to \mathbb{R}^n .

The simplest example of a manifold is, of course, just \mathbb{R}^n itself; for each $x \in \mathbb{R}^n$ we can take U to be all of \mathbb{R}^n . Clearly, \mathbb{R}^n supplied with an equivalent metric (one which makes it homeomorphic to \mathbb{R}^n with the usual metric), is also a manifold. Indeed, a hasty recollection of the definition shows that anything homeomorphic to a manifold is also a manifold—the specific metric with which M is endowed plays almost no role, and we shall almost never mention it.

[If you know anything about topological spaces, you can replace “metric space” by “topological space” in our definition; this new definition allows some pathological creatures which are not metrizable and which fail to have other properties one might carelessly assume must be possessed by spaces which are locally so nice. Appendix A contains remarks, supplementing various chapters, which should be consulted if one allows a manifold to be non-metrizable.]

The second simplest example of a manifold is an open ball in \mathbb{R}^n ; in this case we can take U to be the entire open ball since an open ball in \mathbb{R}^n is homeomorphic to \mathbb{R}^n . This example immediately suggests the next: any open

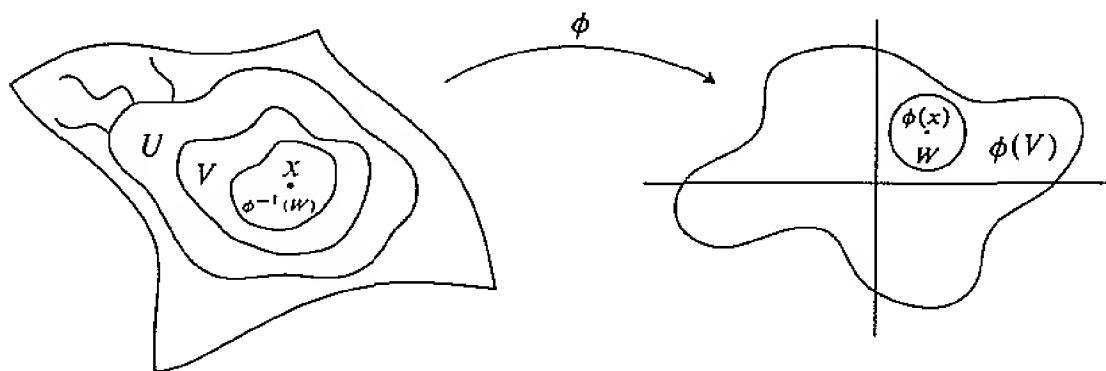
subset V of \mathbb{R}^n is a manifold—for each $x \in V$ we can choose U to be some open ball with $x \in U \subset V$. Exercising a mathematician's penchant for generalization,



we immediately announce a proposition whose proof is left to the reader: An open subset of a manifold is also a manifold (called, quite naturally, an open submanifold of the original manifold).

The open subsets of \mathbb{R}^n already provide many different examples of manifolds (just how many is the subject of Problem 24), though by no means all. Before proceeding to examine other examples, which constitute most of this chapter, some preliminary remarks need to be made.

If x is a point of a manifold M , and U is a neighborhood of x (U contains some open set V with $x \in V$) which is homeomorphic to \mathbb{R}^n by a homeomorphism $\phi: U \rightarrow \mathbb{R}^n$, then $\phi(V) \subset \mathbb{R}^n$ is an open set containing $\phi(x)$. Conse-



quently, there is an open ball W with $\phi(x) \in W \subset \phi(V)$. Thus $x \in \phi^{-1}(W) \subset V \subset U$. Since $\phi: V \rightarrow \mathbb{R}^n$ is continuous, the set $\phi^{-1}(W)$ is open in V , and thus open in M ; it is, of course, homeomorphic to W , and thus to \mathbb{R}^n . This complicated little argument just shows that we can always choose the neighborhood U in our definition to be an open neighborhood.

With a little thought, it begins to appear that, in fact, U *must* be open. But to prove this, we need the following theorem, stated here without proof.*

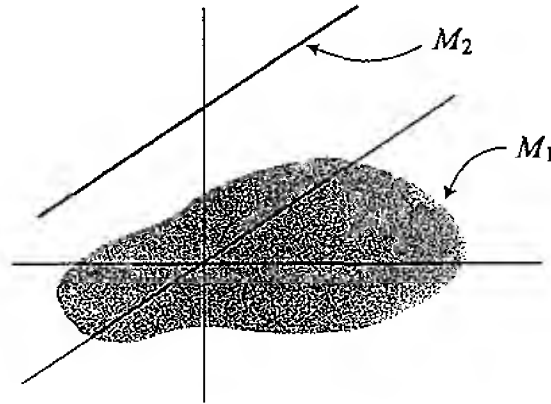
1. THEOREM. If $U \subset \mathbb{R}^n$ is open and $f: U \rightarrow \mathbb{R}^n$ is one-one and continuous, then $f(U) \subset \mathbb{R}^n$ is open. (It follows that $f(V)$ is open for any open $V \subset U$, so f^{-1} is continuous, and f is a homeomorphism.)

Theorem 1 is called “Invariance of Domain”, for it implies that the property of being a “domain” (a connected open set) is invariant under one-one continuous maps into \mathbb{R}^n . The proof that the neighborhood U in our definition must be open is a simple deduction from Invariance of Domain, left to the reader as an easy exercise (it is also easy to see that if Theorem 1 were false, then there would be an example where the U in our definition was not open).

We next turn our attention to the integer n appearing in our definition. Notice that n may depend on the point x . For example, if $M \subset \mathbb{R}^3$ is

$$\begin{aligned} M &= \{(x, y, z) : z = 0\} \cup \{(x, y, z) : x = 0 \text{ and } z = 1\} \\ &= M_1 \cup M_2, \end{aligned}$$

then we can choose $n = 2$ for points in M_1 and $n = 1$ for points in M_2 . This



example, by the way, is an unnecessarily complicated device for producing one manifold from two. In general, given M_1 and M_2 , with metrics d_1 and d_2 , we can first replace each d_i with an equivalent metric \bar{d}_i such that $\bar{d}_i(x, y) < 1$ for all $x, y \in M_i$; for example, we can define

$$\bar{d}_i = \frac{d_i}{1 + d_i} \quad \text{or} \quad \bar{d}_i = \min(d_i, 1).$$

* All proofs require some amount of machinery. The quickest routes are probably provided by Vick, *Homology Theory* and Massey, *Singular Homology Theory*. An old-fashioned, but pleasantly geometric, treatment may be found in Newman, *Topology of Plane Sets*.

Then we can define a metric d on $M = M_1 \cup M_2$ by

$$d(x, y) = \begin{cases} \bar{d}_i(x, y) & \text{if there is some } i \text{ such that } x, y \in M_i \\ 1 & \text{otherwise} \end{cases}$$

(we assume that M_1 and M_2 are disjoint; if not, they can be replaced by new sets which are). In the new space M , both M_1 and M_2 are open sets. If M_1 and M_2 are manifolds, M is clearly a manifold also. This construction can be applied to any number of spaces—even uncountably many; the resulting metric space is called the **disjoint union** of the metric spaces M_i . A disjoint union of manifolds is a manifold. In particular, since a space with one point is a manifold, so is any discrete space M , defined by the metric

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y. \end{cases}$$

Although different n 's may be required at different points of a manifold M , it would seem that only one n can work at a given point $x \in M$. For the proof of this intuitively obvious assertion we have recourse once again to Invariance of Domain. As a first step, we note that \mathbb{R}^n is not homeomorphic to \mathbb{R}^m when $n \neq m$, for if $n > m$, then there is a one-one continuous map from \mathbb{R}^m into a non-open subset of \mathbb{R}^n . The further deduction, that the n of our definition is unique at each $x \in M$, is left to the reader. This unique n is called the **dimension of M at x** . A manifold has **dimension n** or is **n -dimensional** or is an **n -manifold** if it has dimension n at each point. It is convenient to refer to the manifold M as M^n when we want to indicate that M has dimension n .

Consider once more a discrete space, which is a 0-dimensional manifold. The only compact subsets of such a space are finite subsets. Consequently, an uncountable discrete space is not σ -compact (it cannot be written as a countable union of compact subsets). The same phenomenon occurs with higher-dimensional manifolds, as we see by taking a disjoint union of uncountably many manifolds homeomorphic to \mathbb{R}^n . In these examples, however, the manifold is not connected. We will often need to know that this is the only way in which σ -compactness can fail to hold.

2. THEOREM. If X is a connected, locally compact metric space, then X is σ -compact.

PROOF. For each $x \in X$ consider those numbers $r > 0$ such that the closed ball

$$\{y \in X : d(x, y) \leq r\}$$

is a compact set (there is at least one such $r > 0$, since X is locally compact). The set of all such $r > 0$ is an interval. If, for some x , this set includes all $r > 0$, then X is σ -compact, since

$$X = \bigcup_{n=1}^{\infty} \{y \in X : d(x, y) \leq n\}.$$

If not, then for each $x \in X$ define $r(x)$ to be one-half the least upper bound of all such r .

The triangle inequality implies that

$$\{y \in X : d(x_1, y) \leq r\} \subset \{y \in X : d(x_2, y) \leq r + d(x_1, x_2)\},$$

so that

$$\{y \in X : d(x_1, y) \leq r - d(x_1, x_2)\} \subset \{y \in X : d(x_2, y) \leq r\},$$

which implies that

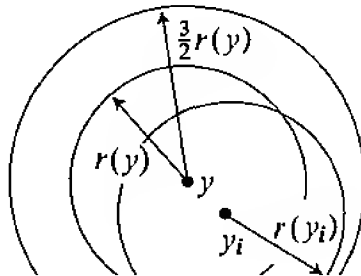
$$(1) \quad r(x_1) \geq r(x_2) - \frac{1}{2}d(x_1, x_2).$$

Interchanging x_1 and x_2 gives

$$(2) \quad |r(x_1) - r(x_2)| \leq \frac{1}{2}d(x_1, x_2),$$

so the function $r: X \rightarrow \mathbb{R}$ is continuous. This has the following important consequence. Suppose $A \subset X$ is compact. Let A' be the union of all closed balls of radius $r(y)$ and center y , for all $y \in A$. Then A' is also compact. The proof is as follows.

Let z_1, z_2, z_3, \dots be a sequence in A' . For each i there is a $y_i \in A$ such that z_i is in the ball of radius $r(y_i)$ with center y_i . Since A is compact, some subsequence of the y_i , which we might as well assume is the sequence itself, converges to some point $y \in A$. Now the closed ball B of radius $\frac{3}{2}r(y)$ and



center y is compact. Since $y_i \rightarrow y$ and since the function r is continuous, eventually the closed balls

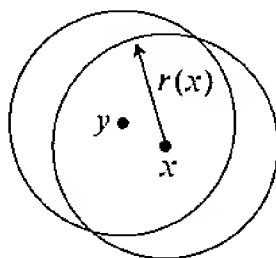
$$\{y \in X : d(y, y_i) \leq r(y_i)\}$$

are contained in B . So the sequence z_i is eventually in the compact set B , and consequently some subsequence converges. Moreover, the limit point is actually in the closed ball of radius $r(y)$ and center y (Problem 10). Thus A' is compact.

Now let $x_0 \in X$ and consider the compact sets

$$\begin{aligned} A_1 &= \{x_0\} \\ A_{n+1} &= A_n'. \end{aligned}$$

Their union A is clearly open. It is also closed. To see this, suppose that x is a point in the closure of A . Then there is some $y \in A$ with $d(x, y) < \frac{2}{3}r(x)$.



By (1),

$$\begin{aligned} r(y) &\geq r(x) - \frac{1}{2}d(x, y) \\ &> r(x) - \frac{1}{2} \cdot \frac{2}{3}r(x) = \frac{2}{3}r(x) \\ &> d(x, y). \end{aligned}$$

This shows that if $y \in A_n$, then $x \in A_n'$, so $x \in A$.

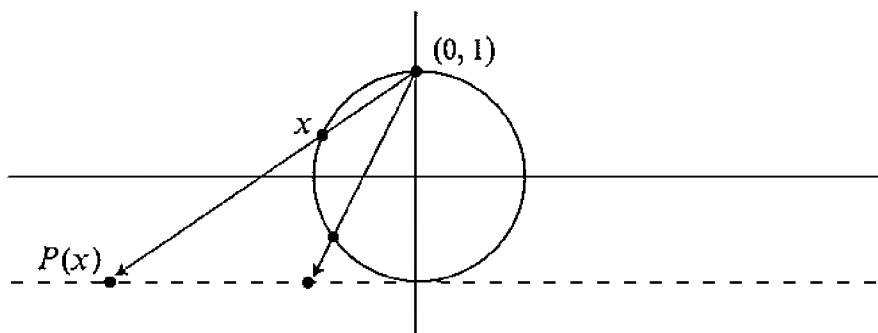
Since X is connected, and $A \neq \emptyset$ is open and closed, it must be that $X = A$, which is σ -compact. ♦

After this hassle with point-set topology, we present the long-promised examples of manifolds. The only connected 1-manifolds are the line \mathbb{R} and the circle, or 1-dimensional sphere, S^1 , defined by

$$S^1 = \{x \in \mathbb{R}^2 : d(x, 0) = 1\}.$$

The function $f: (0, 2\pi) \rightarrow S^1$ defined by $f(\theta) = (\cos \theta, \sin \theta)$ is a homeomorphism; it is even continuous, though not one-one, on $[0, 2\pi]$. We will often denote the point $(\cos \theta, \sin \theta) \in S^1$ simply by $\theta \in [0, 2\pi]$. (Of course, it is always necessary to check that use of this notation is valid.) The function $g: (-\pi, \pi) \rightarrow S^1$, defined by the same formula, is also a homeomorphism; together with f it shows that S^1 is indeed a manifold.

There is another way to prove this, better suited to generalization. The projection P from the point $(0, 1)$ onto the line $\mathbb{R} \times \{-1\} \subset \mathbb{R} \times \mathbb{R}$, illustrated in



the above diagram, is a homeomorphism of $S^1 - \{(0, 1)\}$ onto $\mathbb{R} \times \{-1\}$: this is proved most simply by calculating $P: S^1 - \{(0, 1)\} \rightarrow \mathbb{R} \times \{-1\}$ explicitly. The point $(0, 1)$ may be taken care of similarly, by projecting onto $\mathbb{R} \times \{1\}$, or it suffices to note that S^1 is “homogeneous”—there is a homeomorphism taking any point into any other (namely, an appropriate rotation of \mathbb{R}^2). Considerations similar to these now show that the n -sphere

$$S^n = \{x \in \mathbb{R}^{n+1} : d(x, 0) = 1\}$$

is an n -manifold. The 2-sphere S^2 , commonly known as “the sphere”, is our first example of a compact 2-manifold or surface.

From these few manifolds we can already construct many others by noting that if M_i are manifolds of dimension n_i ($i = 1, 2$), then $M_1 \times M_2$ is an $(n_1 + n_2)$ -manifold. In particular

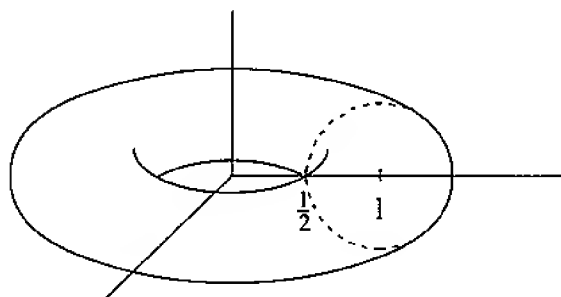
$$\underbrace{S^1 \times \cdots \times S^1}_{n \text{ times}}$$

is called the n -torus, while $S^1 \times S^1$ is commonly called “the torus”. It is obviously homeomorphic to a subset of \mathbb{R}^4 , and it is also homeomorphic to a certain subset of \mathbb{R}^3 which is what most people have in mind when they speak of

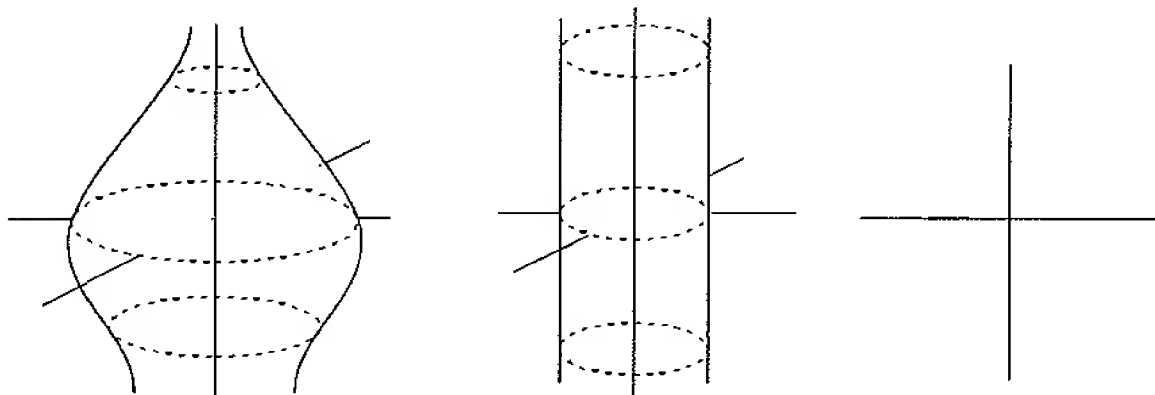
“the torus”: This subset may be obtained by revolving the circle

$$\{(0, y, z) \in \mathbb{R}^3 : (y - 1)^2 + z^2 = 1/4\}$$

around the z -axis. The same construction may be applied to any 1-manifold



contained in $\{(0, y, z) \in \mathbb{R}^3 : y > 0\}$. The resulting surface, called a **surface of revolution**, has components homeomorphic either to the torus or to the cylinder $S^1 \times \mathbb{R}$, the latter of which is also homeomorphic to the annulus, the region of the plane contained between two concentric circles.



The next simplest compact 2-manifold is the 2-holed torus. To provide a more



explicit description of the 2-holed torus, it is easiest to begin with a “handle”, a space homeomorphic to a torus with a hole cut out; more precisely, we throw

away all the points on one side of a certain circle, which remains in our handle, and which will be referred to as the boundary of the handle. The 2-holed

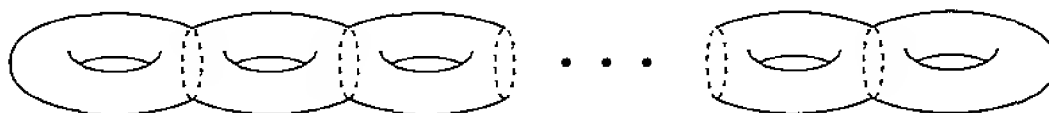


torus may be obtained by piecing two of these together; it is also described as

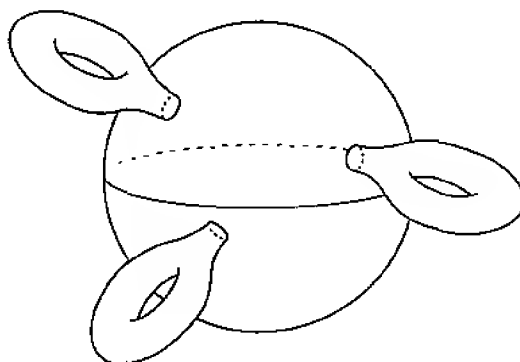


the disjoint union of two handles with corresponding points on the boundaries "identified".

The n -holed torus may be obtained by repeated applications of this procedure. It is homeomorphic to the space obtained by starting with the disjoint

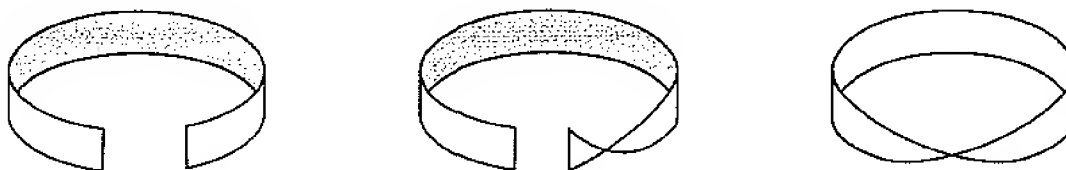


union of n handles and a sphere with n holes, and then identifying points on the boundary of the i^{th} handle with corresponding points on the i^{th} boundary piece of the sphere.



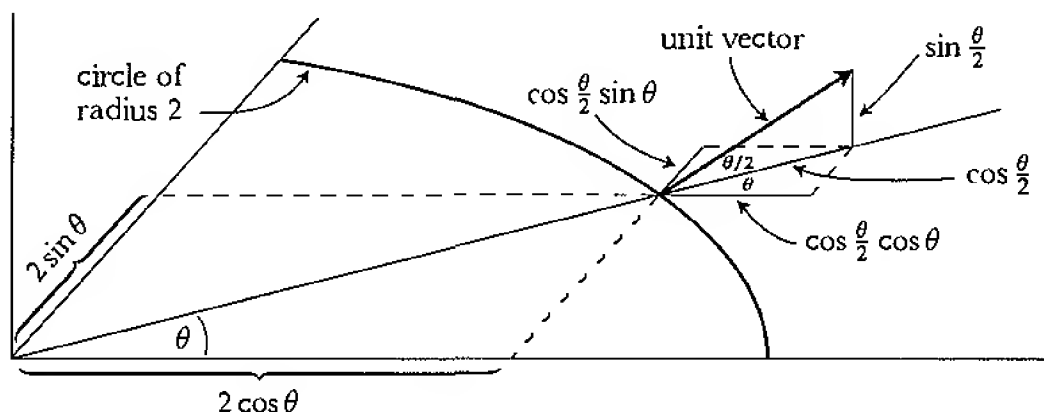
There is one 2-manifold of which most budding mathematicians make the acquaintance when they still know more about paper and paste than about

metric spaces—the famous *Möbius strip*, which you “make” by giving a strip of paper a half twist before pasting its ends together. This can be described

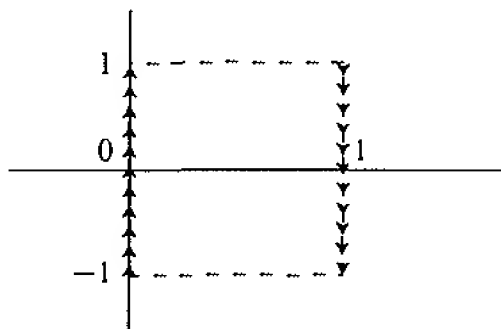


analytically as the image in \mathbb{R}^3 of the function $f : [0, 2\pi] \times (-1, 1) \rightarrow \mathbb{R}^3$ defined by

$$f(\theta, t) = \left(2 \cos \theta + t \cos \frac{\theta}{2} \cos \theta, 2 \sin \theta + t \cos \frac{\theta}{2} \sin \theta, t \sin \frac{\theta}{2} \right).$$

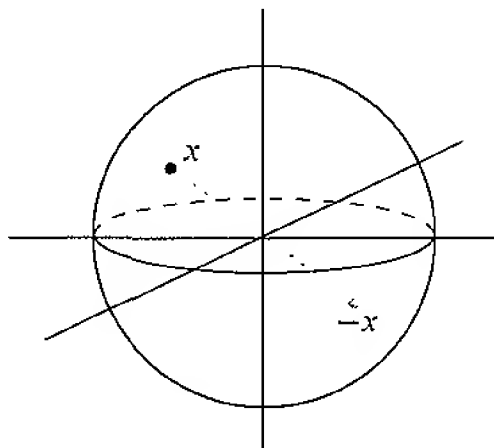


If we define f on $[0, 2\pi] \times [-1, 1]$ instead, we obtain the Möbius strip with a boundary; as investigation of the paper model will show, this boundary is homeomorphic to a circle, not to two disjoint circles. With our recently introduced terminology, the Möbius strip can also be described as $[0, 1] \times (-1, 1)$ with $(0, t)$ and $(1, -t)$ “identified”.

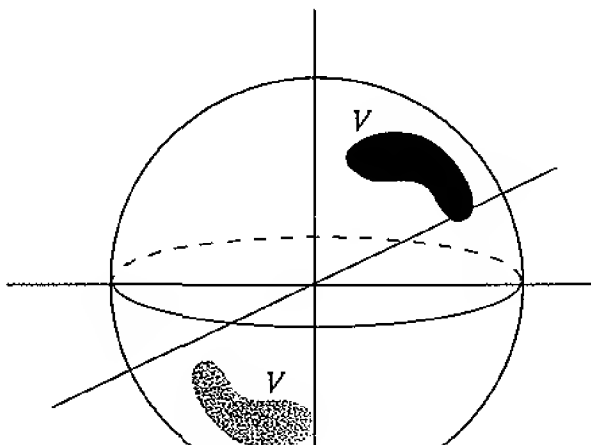


We have not yet had to make precise this notion of “identification”, but our next example will force the issue. We wish to identify each point $x \in S^2$ with

its antipodal point $-x \in S^2$. The space which results, the **projective plane**, \mathbb{P}^2 , is a lot harder to visualize than previous examples; indeed, there is no subset of \mathbb{R}^3 which represents it adequately.



The precise definition of \mathbb{P}^2 uses the same trick that mathematicians always use when they want two things which are not equal to be equal. The *points* of \mathbb{P}^2 are defined to be the sets $\{p, -p\}$ for $p \in S^2$. We will denote this set by $[p] \in \mathbb{P}^2$, so that $[-p] = [p]$. We thus have a map $f: S^2 \rightarrow \mathbb{P}^2$ given by $f(p) = [p]$, for which $f(p) = f(q)$ implies $p = \pm q$. We will postpone for a while the problem of defining the metric giving the distance between two points $[p]$ and $[q]$, but we can easily say what the open sets will turn out to be (and this is all you need to know in order to check that \mathbb{P}^2 is a surface). A subset $U \subset \mathbb{P}^2$ will be open if and only if $f^{-1}(U) \subset S^2$ is open. This just means that the open sets of \mathbb{P}^2 are of the form $f(V)$ where $V \subset S^2$ is an open set with the additional important property that if it contains p it also contains $-p$.



In exactly the same way, we could have defined the points of the Möbius strip M to be

$$\text{all points } (s, t) \in (0, 1) \times (-1, 1)$$

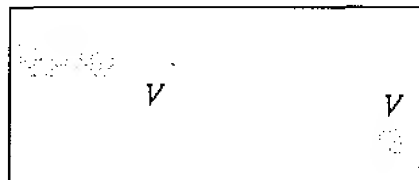
together with

$$\text{all sets } \{(0, t), (1, -t)\}, \quad \text{denoted by } [(0, t)] \text{ or } [(1, -t)].$$

There is a map $f: [0, 1] \times (-1, 1) \rightarrow M$ given by

$$f((s, t)) = \begin{cases} (s, t) & \text{if } s \neq 0, 1 \\ [(s, t)] & \text{if } s = 0 \text{ or } 1, \end{cases}$$

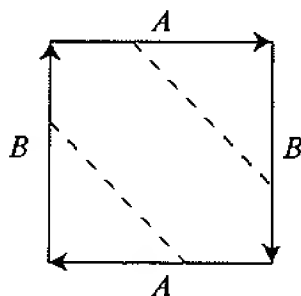
and $U \subset M$ is open if and only if $f^{-1}(U) \subset [0, 1] \times (-1, 1)$ is open, so that the open sets of M are of the form $f(V)$ where V is open and contains $(s, -t)$ whenever it contains (s, t) for $s = 0$ or 1 .



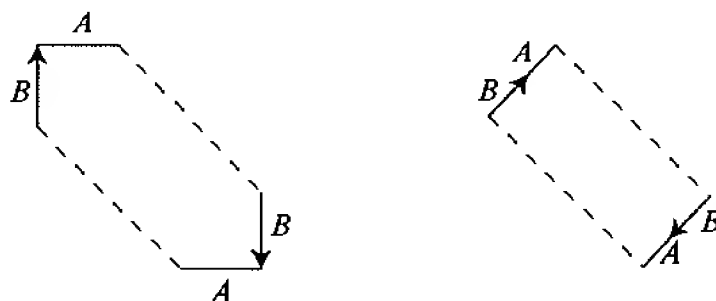
To get an idea of what \mathbb{P}^2 looks like, we can make things easier for ourselves by first throwing away all points of S^2 below the (x, y) -plane, since they are identified with points above the (x, y) -plane anyway. This leaves the upper hemisphere (including the bounding circle), which is homeomorphic to the disc

$$D^2 = \{x \in \mathbb{R}^2 : d(x, 0) \leq 1\},$$

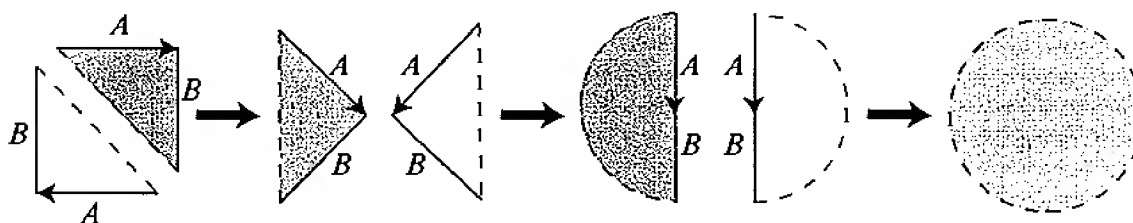
and we must identify each $p \in S^1$ with $-p \in S^1$. Squaring things off a bit, this is the same as identifying points on the sides of a square according to the scheme shown below (points on sides with the same label are identified in such a way that the heads of the arrows are identified with each other). The dotted lines in this picture are the key to understanding \mathbb{P}^2 . If we distort the



region between them a bit we see that the front part of B followed by the back part of A , at the upper left, is to be identified with the same thing at the lower right, in reverse direction; in other words, we obtain a Möbius strip with

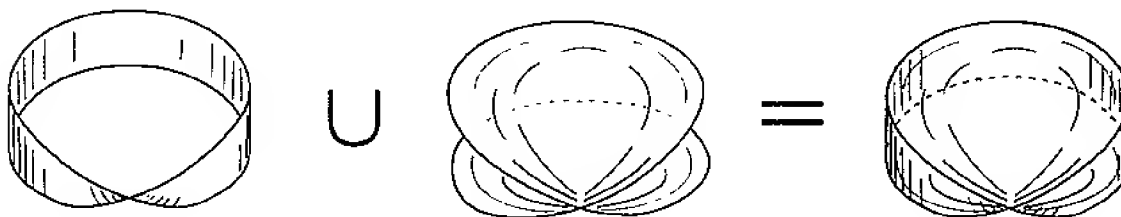


a boundary (namely, the dotted line, which is a single circle). If this Möbius strip is removed, we are left with two pieces which can be rearranged to form something homeomorphic to a disc. The projective plane is thus obtained from



the disjoint union of a disc and a Möbius strip with a boundary, by identifying points on the boundary and points on the boundary of the disc, both of which are circles. Thus to make a model of \mathbb{P}^2 we just have to sew a circular piece of cloth and a cloth Möbius strip together along their edges. Unfortunately, a little experimentation will convince you that this cannot be done (without having the two pieces of cloth pass through each other).

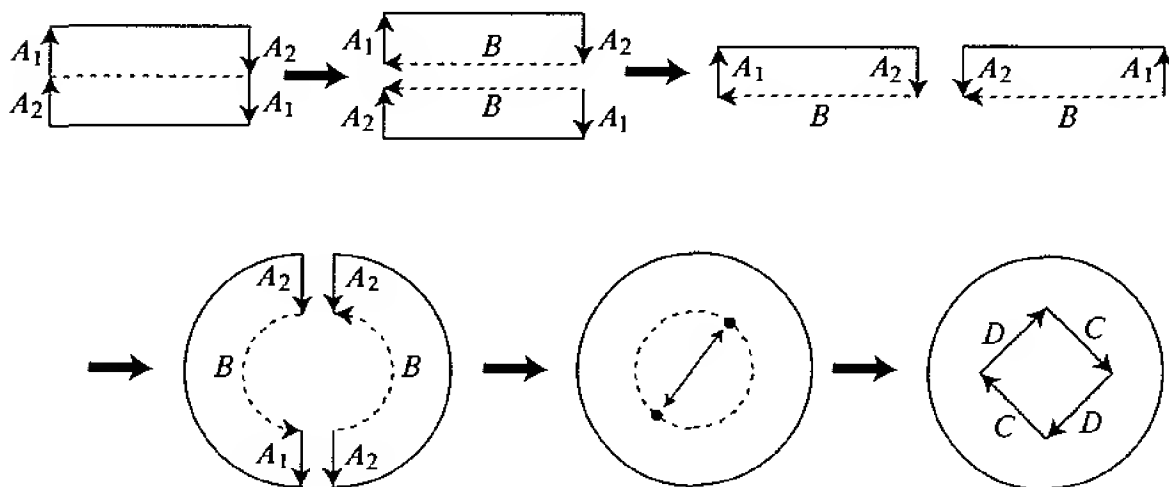
The subset of \mathbb{R}^3 obtained as the union of the Möbius strip and a disc, although not homeomorphic to \mathbb{P}^2 , can still be described mathematically in terms



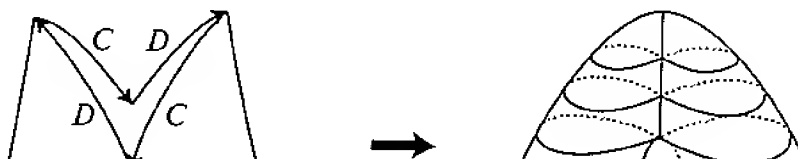
of \mathbb{P}^2 . There is clearly a continuous function $f: \mathbb{P}^2 \rightarrow \mathbb{R}^3$ whose image is this subset; moreover, although f is not one-one, it is **locally one-one**, that is, every point $p \in \mathbb{P}^2$ has a neighborhood U on which f is one-one. Such a function f

is called a **topological immersion** (the single word “immersion” has a more specialized meaning, explained in Chapter 2). We can thus say that \mathbb{P}^2 can be topologically immersed in \mathbb{R}^3 , although not topologically imbedded (there is no homeomorphism f from \mathbb{P}^2 to a subset of \mathbb{R}^3). In \mathbb{R}^4 , however, with an extra dimension to play around with, the disc can be added so as not to intersect the Möbius strip.

Another topological immersion of \mathbb{P}^2 in \mathbb{R}^3 can be obtained by first immersing the Möbius strip so that its boundary circle lies in a plane; this can be done in the following way. The figures below show that the Möbius strip may be obtained from an annulus by identifying opposite points of the inner circle. (This is also obvious from the fact that the Möbius strip is the projective plane with a disc removed.) This inner circle can be replaced by a quadrilateral. When the



resulting figure is drawn up into 3-space and the appropriate identifications are made we obtain the “cross-cap”. The cross-cap together with the disc at the bottom is a topologically immersed \mathbb{P}^2 .



The one gap in the preceding discussion is the definition of a metric for \mathbb{P}^2 . The missing metric can be supplied by an appeal to Problem 3-1, which will later be used quite often, and which the reader should peruse sometime before reading Chapter 3. Roughly speaking, it shows that things like \mathbb{P}^2 , which ought to be manifolds, are. (Those who know about topological spaces will recognize it as a disguised case of the Urysohn Metrization Theorem.) For the present, however, we will obtain our metric by a trick that simultaneously provides an imbedding of \mathbb{P}^2 in \mathbb{R}^4 . Consider the function $f: S^2 \rightarrow \mathbb{R}^4$ defined by

$$f(x, y, z) = (yz, xz, xy, x^2 + 2y^2 + 3z^2).$$

(Clearly $f(p) = f(-p)$.) We maintain that $f(p) = f(q)$ implies that $p = \pm q$. To prove this, suppose that $f(x, y, z) = f(a, b, c)$. We have, first of all

$$(1) \quad \begin{aligned} yz &= bc \\ xz &= ac \\ xy &= ab. \end{aligned}$$

If $a, b, c \neq 0$, this leads to

$$(2) \quad \begin{aligned} y &= \frac{bx}{a} \\ z &= \frac{cx}{a}. \end{aligned}$$

Now

$$\begin{aligned} (x + y + z)^2 &= x^2 + y^2 + z^2 + 2(xy + xz + yz) \\ &= 1 + 2(xy + xz + yz), \end{aligned}$$

so we also have

$$(x + y + z)^2 = (a + b + c)^2,$$

hence

$$(3) \quad a + b + c = \pm(x + y + z).$$

Using (2), this gives

$$a + b + c = \pm x \left(1 + \frac{b}{a} + \frac{c}{a} \right) = \pm x \left(\frac{a + b + c}{a} \right),$$

so $x = \pm a$. Similarly, we obtain $y = \pm b$, $z = \pm b$, with the same sign (which comes from (3)) holding for all three equations. In this case we have proved our contention without even using the fourth coordinate of f . Now suppose $a = 0$. If $x \neq 0$, then (1) would immediately give $y = z = 0$, so that

$$(x, y, z) = (\pm 1, 0, 0).$$

But $y = z = 0$ implies (by (1) again) that $bc = 0$, so $b = 0$ or $c = 0$ and

$$(a, b, c) = (0, \pm 1, 0) \text{ or } (0, 0, \pm 1).$$

These equations clearly contradict

$$x^2 + 2y^2 + 3z^2 = a^2 + 2b^2 + 3c^2.$$

Thus $x = 0$ also, and we have

$$(4) \quad yz = bc$$

$$(5) \quad 2y^2 + 3z^2 = 2b^2 + 3c^2$$

$$(6) \quad y^2 + z^2 = 1$$

$$b^2 + c^2 = 1.$$

But (6) implies that

$$\begin{aligned} 2y^2 + 3z^2 &= 2y^2 + 3(1 - y^2) \\ &= 3 - y^2, \end{aligned}$$

and similarly for b and c , so (5) gives

$$(7) \quad \begin{aligned} 3 - y^2 &= 3 - b^2 \\ y &= \pm b. \end{aligned}$$

Now (4) gives

$$(8) \quad z = \pm c$$

(this holds even if $y = b = 0$, since then $z, c = \pm 1$). Clearly, (4) also shows that the same sign holds in (7) and (8), which completes the proof.

Since $f(p) = f(q)$ precisely when $p = \pm q$, we can define $\tilde{f}: \mathbb{P}^2 \rightarrow \mathbb{R}^4$ by

$$\tilde{f}([p]) = f(p).$$

This map is one-one and we can use it to define the metric in \mathbb{P}^2 :

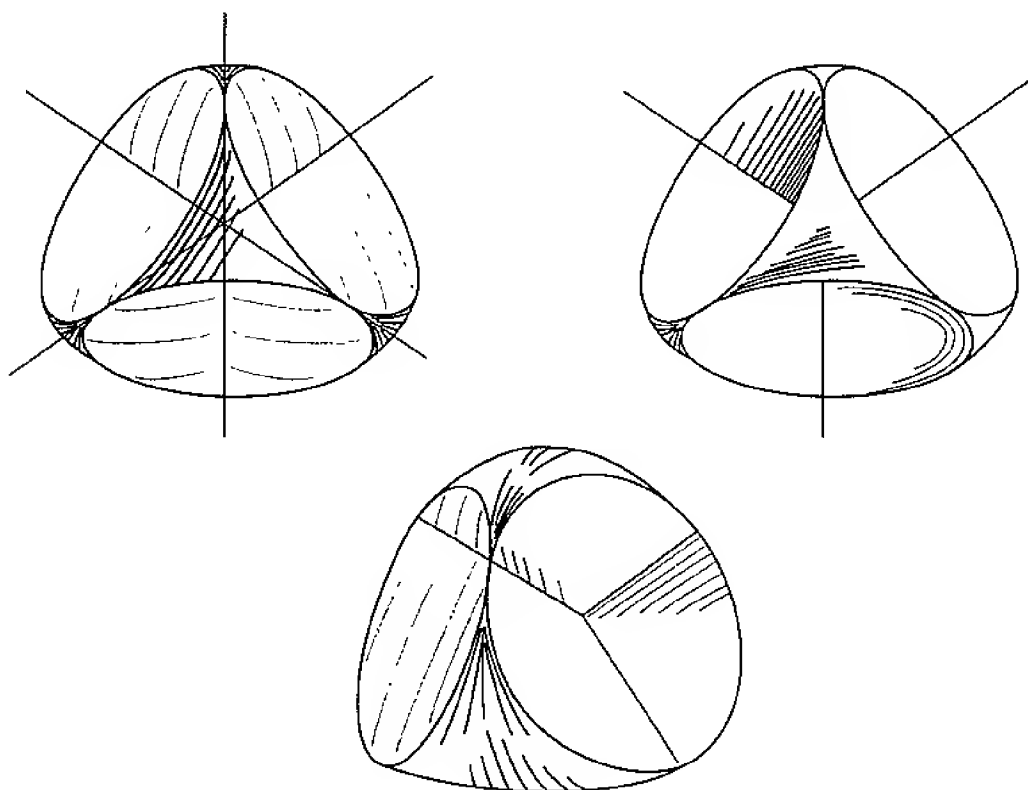
$$\tilde{d}([p], [q]) = d(\tilde{f}([p]), \tilde{f}([q])) = d(f(p), f(q)).$$

Then one can check that the open sets are indeed the ones described above.

By the way, the map $g: \mathbb{P}^2 \rightarrow \mathbb{R}^3$ defined by the first 3 components of f ,

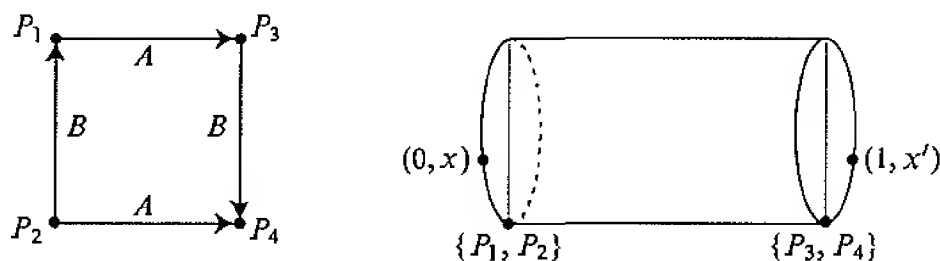
$$g([x, y, z]) = (yz, xz, xy)$$

is a topological immersion of \mathbb{P}^2 in \mathbb{R}^3 . The image in \mathbb{R}^3 is Steiner's "Roman surface".

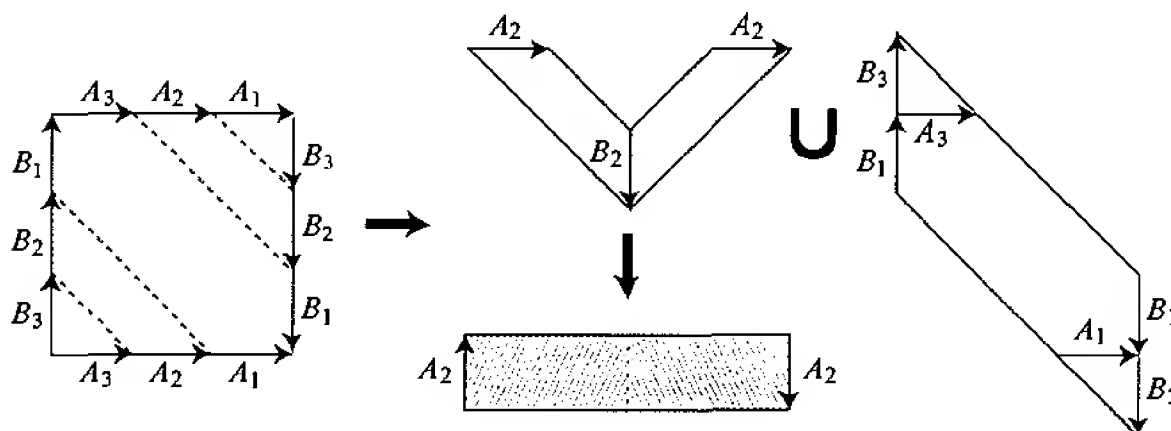


With the new surface \mathbb{P}^2 at our disposal, we can create other surfaces in the same way as the n -holed torus. For example, to a handle we can attach a projective space with a hole cut out, or, what amounts to the same thing, a Möbius strip. The closest we can come to picturing this is by drawing a cross-cap sticking on a torus. We can also join together a pair of projective planes with holes cut out, which amounts to sewing two Möbius strips together along their boundary. Although this can be pictured as two cross-caps joined together, it has a nicer, and famous, representation. Consider the surface obtained from the square with identifications indicated below; it may also be obtained from the cylinder $[0, 1] \times S^1$ by identifying $(0, x) \in [0, 1] \times S^1$ with $(1, x')$, where x' is the reflection of x through a fixed diameter of the circle. Notice that the

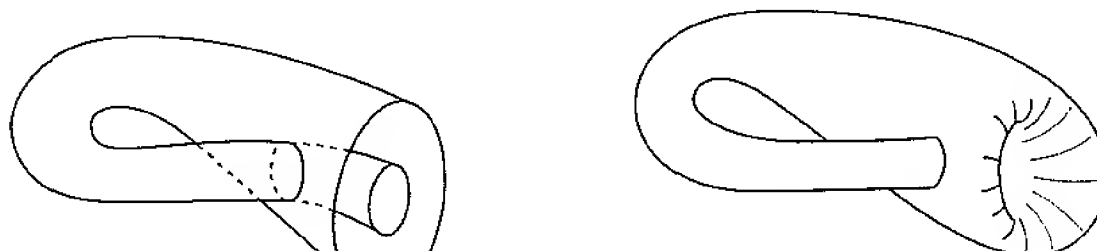
identifications on the square force P_1 , P_2 , P_3 , and P_4 to be identified, so that the set $\{P_1, P_2, P_3, P_4\}$ is a single point of our new space. The dotted lines



below, dividing the sides into thirds, form a single circle, which separates the surface into two parts, one of which is shaded.



Rearrangement of the two parts shows that this surface is precisely two Möbius strips with corresponding points on their boundary identified. The description in terms of $[0, 1] \times S^1$ immediately suggests an immersion of the surface. Turning one end of the cylinder around and pushing it through itself orients the left-hand boundary so that $(0, x)$ is directly opposite $(1, x')$, to which it can then be joined, forming the "Klein bottle".



Examples of higher-dimensional manifolds will not be treated in nearly such detail, but, in addition to the family of n -manifolds S^n , we will mention the related family of “projective spaces”. **Projective n -space** \mathbb{P}^n is defined as the collection of all sets $\{p, -p\}$ for $p \in S^n$. The description of the open sets in \mathbb{P}^n is precisely analogous to the description for \mathbb{P}^2 . Although these spaces seem to form a family as regular as the family S^n , we will see later that the spaces \mathbb{P}^n for even n differ in a very important way from the same spaces for odd n .

One further definition is needed to complete this introduction to manifolds. We have already discussed some spaces which are not manifolds only because they have a “boundary”, for example, the Möbius strip and the disc. Points on these “boundaries” do not have neighborhoods homeomorphic to \mathbb{R}^n , but they do have neighborhoods homeomorphic to an important subset of \mathbb{R}^n . The (closed) **half-space** \mathbb{H}^n is defined by

$$\mathbb{H}^n = \{(x^1, \dots, x^n) \in \mathbb{R}^n : x^n \geq 0\}.$$

A **manifold-with-boundary** is a metric space M with the following property:

If $x \in M$, then there is some neighborhood U of x and some integer $n \geq 0$ such that U is homeomorphic to either \mathbb{R}^n or \mathbb{H}^n .

A point in a manifold-with-boundary cannot have a neighborhood homeomorphic to both \mathbb{R}^n and \mathbb{H}^n (Invariance of Domain again); we can therefore distinguish those points $x \in M$ having a neighborhood homeomorphic to \mathbb{H}^n . The set of all such x is called the **boundary** of M and is denoted by ∂M . If M is actually a manifold, then $\partial M = \emptyset$. Notice that if M is a subset of \mathbb{R}^n , then ∂M is not necessarily the same as the boundary of M in the old sense (defined for any subset of \mathbb{R}^n); indeed, if M is a manifold-with-boundary of dimension $< n$, then all points of M will be boundary points of M .

If manifolds-with-boundary are studied as frequently as manifolds, it becomes bothersome to use this long designation. Often, the word “manifold” is used for “manifold-with-boundary”. A manifold in our sense is then called “non-bounded”; a non-bounded compact manifold is called a “closed manifold”. We will stick to the other terminology, but will sometimes use “bounded manifold” instead of “manifold-with-boundary”.

PROBLEMS

1. Show that if d is a metric on X , then both $\bar{d} = d/(1 + d)$ and $\bar{d} = \min(1, d)$ are also metrics and that they are equivalent to d (i.e., the identity map $1: (X, d) \rightarrow (X, \bar{d})$ is a homeomorphism).
2. If (X_i, d_i) are metric spaces, for $i \in I$, with metrics $d_i < 1$, and $X_i \cap X_j = \emptyset$ for $i \neq j$, then (X, d) is a metric space, where $X = \bigcup_i X_i$, and $d(x, y) = d_i(x, y)$ if $x, y \in X_i$ for some i , while $d(x, y) = 1$ otherwise. Each X_i is an open subset of X , and Y is homeomorphic to X if and only if $Y = \bigcup_i Y_i$ where the Y_i are disjoint open sets and Y_i is homeomorphic to X_i for each i . The space (X, d) (or any space homeomorphic to it) is called the **disjoint union** of the spaces X_i .
3. (a) Every manifold is locally compact.
 (b) Every manifold is locally pathwise connected, and a connected manifold is pathwise connected.
 (c) A connected manifold is arcwise connected. (A path is a continuous image of $[0, 1]$, but an arc is a *one-one* continuous image. A difficult theorem states that every path contains an arc between its end points, but a direct proof of arcwise-connectedness can be given for manifolds.)
4. A space X is called locally connected if for each $x \in X$ it is the case that every neighborhood of x contains a connected neighborhood.
 (a) Connectedness does not imply local connectedness.
 (b) An open subset of a locally connected space is locally connected.
 (c) X is locally connected if and only if components of open sets are open, so every neighborhood of a point in a locally connected space contains an *open* connected neighborhood.
 (d) A locally connected space is homeomorphic to the disjoint union of its components.
 (e) Every manifold is locally connected, and consequently homeomorphic to the disjoint union of its components, which are open submanifolds.
5. (a) The neighborhood U in our definition of a manifold is always open.
 (b) The integer n in our definition is unique for each x .
6. (a) A subset of an n -manifold is an n -manifold if and only if it is open.
 (b) If M is connected, then the dimension of M at x is the same for all $x \in M$.
7. (a) If $U \subset \mathbb{R}$ is an interval and $f: U \rightarrow \mathbb{R}$ is continuous and one-one, then f is either increasing or decreasing.

- (b) The image $f(U)$ is open.
- (c) The map f is a homeomorphism.

8. For this problem, assume

- (1) (The Generalized Jordan Curve Theorem) If $A \subset \mathbb{R}^n$ is homeomorphic to S^{n-1} , then $\mathbb{R}^n - A$ has 2 components, and A is the boundary of each.
- (2) If $B \subset \mathbb{R}^n$ is homeomorphic to $D^n = \{x \in \mathbb{R}^n : d(x, 0) \leq 1\}$, then $\mathbb{R}^n - B$ is connected.

- (a) One component of $\mathbb{R}^n - A$ (the “outside of A ”) is unbounded, and the other (the “inside of A ”) is bounded.
- (b) If $U \subset \mathbb{R}^n$ is open, $A \subset U$ is homeomorphic to S^{n-1} and $f: U \rightarrow \mathbb{R}^n$ is one-one and continuous (so that f is a homeomorphism on A), then $f(\text{inside of } A) = \text{inside of } f(A)$. (First prove \subset .)
- (c) Prove Invariance of Domain.

- 9. (a) Give an elementary proof that \mathbb{R}^1 is not homeomorphic to \mathbb{R}^n for $n > 1$.
- (b) Prove directly from the Generalized Jordan Curve Theorem that \mathbb{R}^m is not homeomorphic to \mathbb{R}^n for $m \neq n$.

10. In the proof of Theorem 2, show that the limit of a convergent subsequence of the z_i is actually in the closed ball of radius $r(y)$ and center y .

11. Every connected manifold (which is a metric space) has a countable base for its topology, and a countable dense subset.

- 12. (a) Compute the composition $f = S^1 - \{(0, 1)\} \xrightarrow{P} \mathbb{R}^1 \times \{-1\} \rightarrow \mathbb{R}^1$ explicitly for the map P on page 7, and show that it is a homeomorphism.
- (b) Do the same for $f: S^{n-1} - \{(0, \dots, 0, 1)\} \rightarrow \mathbb{R}^{n-1}$.

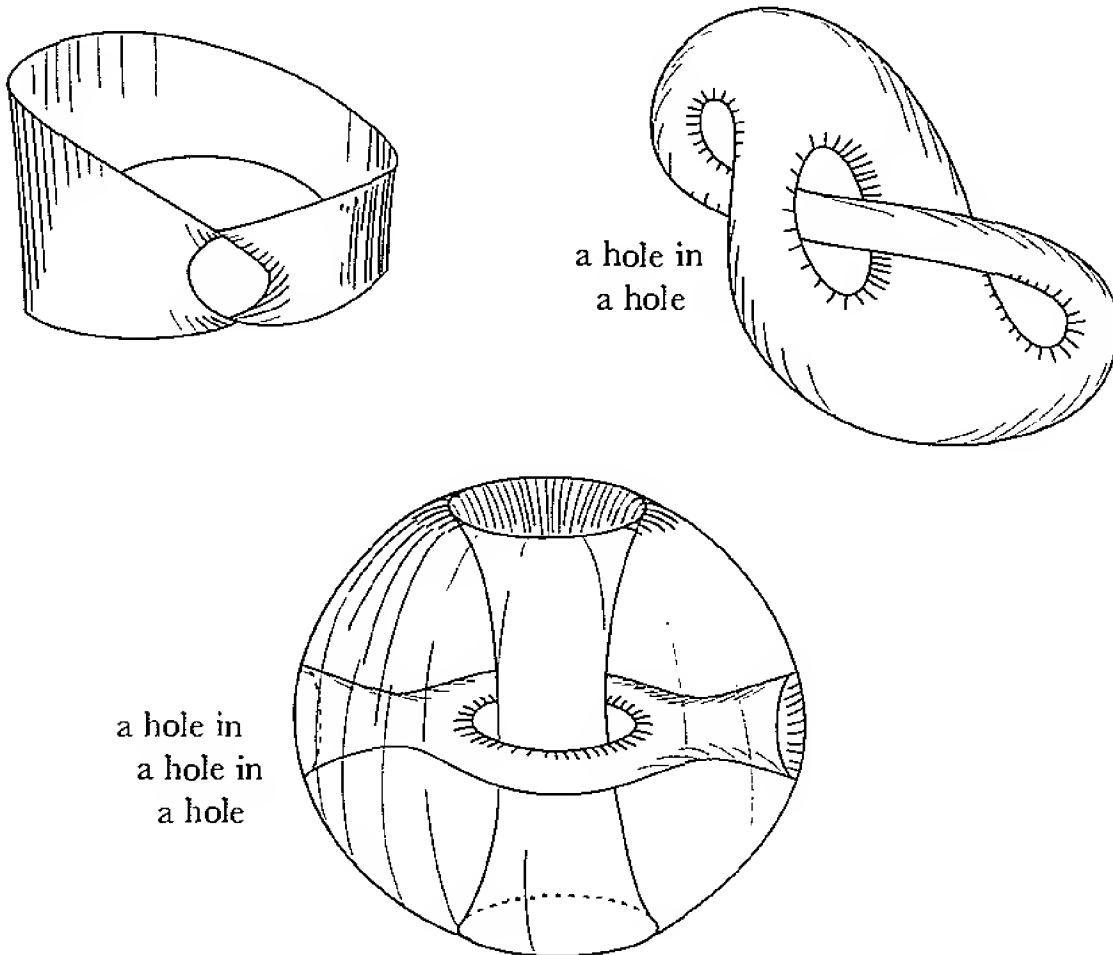
- 13. (a) The text describes the open subsets of \mathbb{P}^2 as sets of the form $f(V)$, where $V \subset S^2$ is open and contains $-p$ whenever it contains p . Show that this last condition is actually unnecessary.
- (b) The analogous condition is necessary for the Möbius strip, which is discussed immediately afterwards. Explain how the two cases differ.

- 14. (a) Check that the metric defined for \mathbb{P}^2 gives the open sets described in the text.
- (b) Check that \mathbb{P}^2 is a surface.

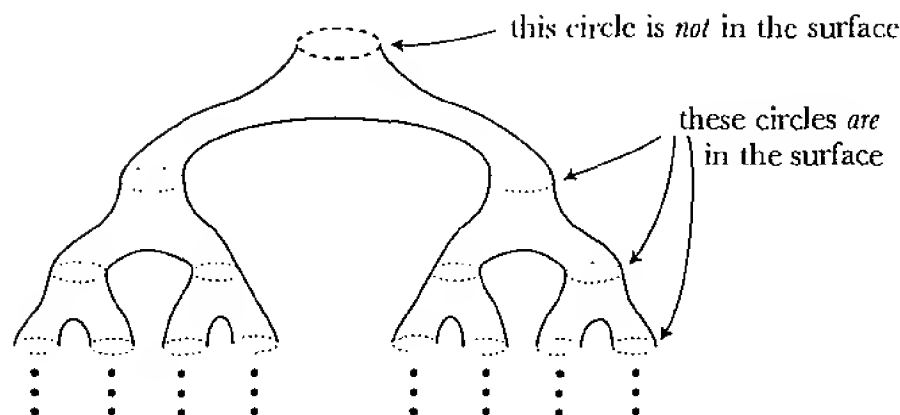
15. (a) Show that \mathbb{P}^1 is homeomorphic to S^1 .

(b) Since we can consider $S^{n-1} \subset S^n$, and since antipodal points in S^{n-1} are still antipodal when considered as points in S^n , we can consider $\mathbb{P}^{n-1} \subset \mathbb{P}^n$ in an obvious way. Show that $\mathbb{P}^n - \mathbb{P}^{n-1}$ is homeomorphic to interior $D^n = \{x \in \mathbb{R}^n : d(x, 0) < 1\}$.

16. A classical theorem of topology states that every compact surface other than S^2 is obtained by gluing together a certain number of tori and projective spaces, and that all compact surfaces-with-boundary are obtained from these by cutting out a finite number of discs. To which of these “standard” surfaces are the following homeomorphic?



17. Let $C \subset \mathbb{R} \subset \mathbb{R}^2$ be the Cantor set. Show that $\mathbb{R}^2 - C$ is homeomorphic to the surface shown at the top of the next page.



18. A locally compact (but non-compact) space X “has one end” if for every compact $C \subset X$ there is a compact K such that $C \subset K \subset X$ and $X - K$ is connected.

(a) \mathbb{R}^n has one end if $n > 1$, but not if $n = 1$.

(b) $\mathbb{R}^n - \{0\}$ does not “have one end” so $\mathbb{R}^n - \{0\}$ is not homeomorphic to \mathbb{R}^m .

19. This problem is a sequel to the previous one; it will be used in Problem 24. An **end** of X is a function ε which assigns to each compact subset $C \subset X$ a non-empty component $\varepsilon(C)$ of $X - C$, in such a way that $C_1 \subset C_2$ implies $\varepsilon(C_2) \subset \varepsilon(C_1)$.

(a) If $C \subset \mathbb{R}$ is compact, then $\mathbb{R} - C$ has exactly 2 unbounded components, the “left” component containing all numbers $<$ some N , the “right” one containing all numbers $>$ some N . If ε is an end of \mathbb{R} , show that $\varepsilon(C)$ is either always the “left” component of $\mathbb{R} - C$, or always the “right” one. Thus \mathbb{R} has 2 ends.

(b) Show that \mathbb{R}^n has only one end ε for $n > 1$. More generally, X has exactly one end ε if and only if X “has one end” in the sense of Problem 18.


(c) This part requires some knowledge of topological spaces. Let $\mathcal{E}(X)$ be the set of all ends of a connected, locally connected, locally compact Hausdorff space X . Define a topology on $X \cup \mathcal{E}(X)$ by choosing as neighborhoods $N_C(\varepsilon_0)$ of an end ε_0 the sets

$$N_C(\varepsilon_0) = \varepsilon_0(C) \cup \{\text{ends } \varepsilon : \varepsilon(C) = \varepsilon_0(C)\},$$

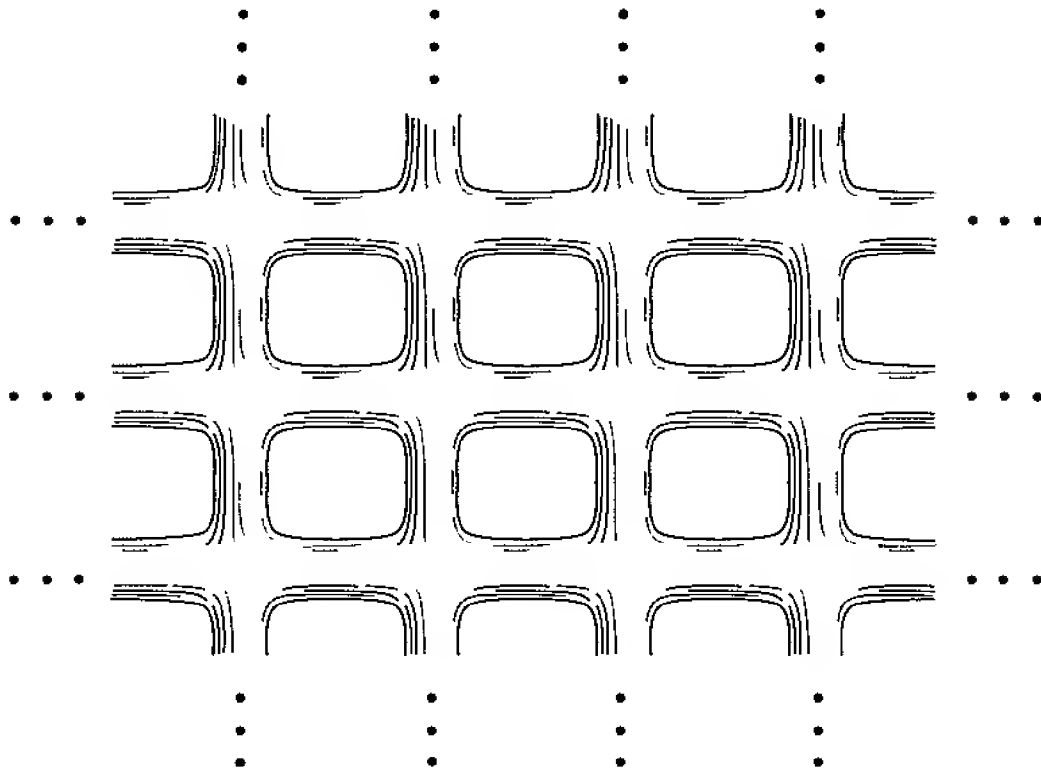
for all compact C . Show that $X \cup \mathcal{E}(X)$ is a compact Hausdorff space. What is $\mathbb{R} \cup \mathcal{E}(\mathbb{R})$, and $\mathbb{R}^n \cup \mathcal{E}(\mathbb{R}^n)$ for $n > 1$?

20. Consider the following three surfaces.

(A) The infinite-holed torus: 

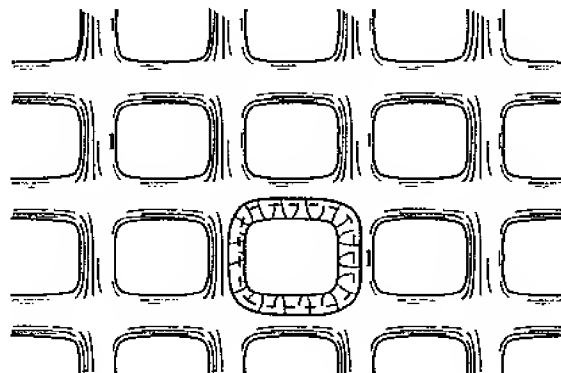
(B) The doubly infinite-holed torus: \cdots  \cdots

(C) The infinite jail cell window:

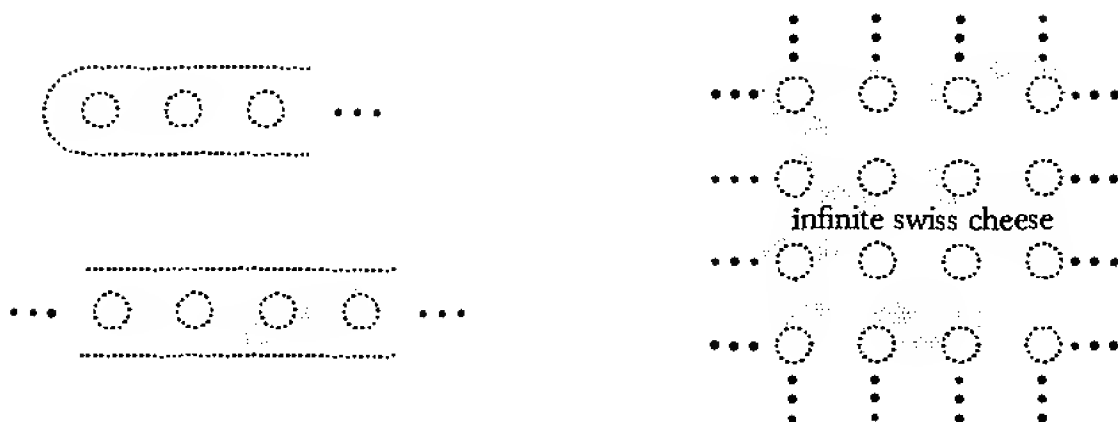


(a) Surfaces (A) and (C) have one end, while surface (B) does not.

(b) Surfaces (A) and (C) are homeomorphic! *Hint:* The region cut out by the lines in the picture below is a cylinder, which occurs at the left of (A). Now draw in two more lines enclosing more holes, and consider the region between the two pairs.



21. (a) The three open subsets of \mathbb{R}^2 shown below are homeomorphic.



(b) The points *inside* the three surfaces of Problem 20 are homeomorphic.

22. (a) Every open subset of \mathbb{R} is homeomorphic to the disjoint union of intervals.

(b) There are only countably many non-homeomorphic open subsets of \mathbb{R} .

23. For the purposes of this problem we will use a consequence of the Urysohn Metrization Theorem, that for any connected manifold M , there is a homeomorphism f from M to a subset of the countable product $\mathbb{R} \times \mathbb{R} \times \cdots$.

(a) If M is a connected non-compact manifold, then there is a continuous function $f: M \rightarrow \mathbb{R}$ such that f "goes to ∞ at ∞ ", i.e., if $\{x_n\}$ is a sequence which is eventually in the complement of every compact set, then $f(x_n) \rightarrow \infty$. (Compare with Problem 2-30.)

(b) Given a homeomorphism $f: M \rightarrow \mathbb{R} \times \mathbb{R} \times \cdots$ and a $g: M \rightarrow \mathbb{R}$ which goes to ∞ at ∞ , define $\bar{f}: M \rightarrow \mathbb{R} \times (\mathbb{R} \times \mathbb{R} \times \cdots)$ by $\bar{f}(x) = (g(x), f(x))$. Show that $\bar{f}(M)$ is closed.

(c) There are at most c non-homeomorphic connected manifolds (where $c = 2^{\aleph_0}$ is the cardinality of \mathbb{R}).

24. (a) It is possible for $\mathbb{R}^2 - A$ and $\mathbb{R}^2 - B$ to be homeomorphic even though A and B are non-homeomorphic closed subsets.

(b) If $A \subset \mathbb{R}^2$ is closed and totally disconnected (the only components of A are points), then $\mathcal{E}(\mathbb{R}^2 - A)$ is homeomorphic to A . Hence $\mathbb{R}^2 - A$ and $\mathbb{R}^2 - B$ are non-homeomorphic if A and B are non-homeomorphic closed totally disconnected sets.

(c) The derived set A' of A is the set of all non-isolated points. We define $A^{(n)}$ inductively by $A^{(1)} = A'$ and $A^{(n+1)} = (A^{(n)})'$. For each n there is a subset A_n of \mathbb{R} such that $A_n^{(n)}$ consists of one point.

*(d) There are \mathfrak{c} non-homeomorphic closed totally disconnected subsets of \mathbb{R}^2 .

Hint: Let C be the Cantor set, and $c_1 < c_2 < c_3 < \dots$ a sequence of points in C . For each sequence $n_1 < n_2 < \dots$, one can add a set A_{n_i} such that its n_i^{th} derived set is $\{c_i\}$.

(e) There are \mathfrak{c} non-homeomorphic connected open subsets of \mathbb{R}^2 .

25. (a) A manifold-with-boundary could be defined as a metric space M with the property that for each $x \in M$ there is a neighborhood U of x and an integer $n \geq 0$ such that U is homeomorphic to an open subset of \mathbb{H}^n .

(b) If M is a manifold-with-boundary, then ∂M is a closed subset of M and ∂M and $M - \partial M$ are manifolds.

(c) If $C_i, i \in I$ are the components of ∂M , and $I' \subset I$, then $M - \bigcup_{i \in I'} C_i$ is a manifold-with-boundary.

26. If $M \subset \mathbb{R}^n$ is a closed set and an n -dimensional manifold-with-boundary, then the topological boundary of M , as a subset of \mathbb{R}^n , is ∂M . This is not necessarily true if M is not a closed subset.

27. (a) Every point (a, b, c) on Steiner's surface satisfies $b^2c^2 + a^2c^2 + a^2b^2 = abc$.

(b) If (a, b, c) satisfies this equation and $0 \neq D = \sqrt{b^2c^2 + a^2c^2 + a^2b^2}$, then (a, b, c) is on Steiner's surface. *Hint:* Let $x = bc/D$, etc.

(c) The set $\{(a, b, c) \in \mathbb{R}^3 : b^2c^2 + a^2c^2 + a^2b^2 = abc\}$ is the union of the Steiner surface and of the portions $(-\infty, -1/2)$ and $(1/2, \infty)$ of each axis.

CHAPTER 2

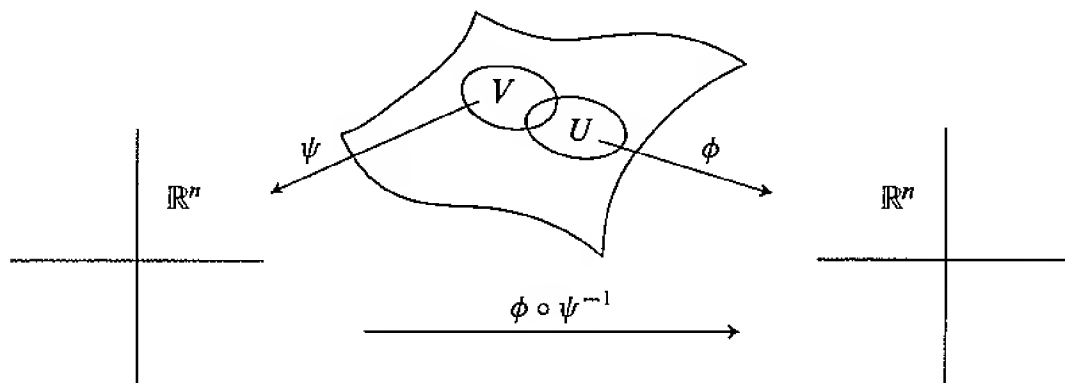
DIFFERENTIABLE STRUCTURES

We are now ready to apply analysis to the study of manifolds. The necessary tools of “advanced calculus”, which the reader should bring along freshly sharpened, are contained in Chapters 2 and 3 of *Calculus on Manifolds*. We will use freely the notation and results of these chapters, *including* some problems, notably 2-9, 2-15, 2-25, 2-26, 2-29, 3-32, and 3-35; however, we will denote the identity map from \mathbb{R}^n to \mathbb{R}^n by I , rather than by π (which will be used often enough in other contexts), so that $I^i(x) = x^i$.

On a general manifold M the notion of a continuous function $f: M \rightarrow \mathbb{R}$ makes sense, but the notion of a differentiable function $f: M \rightarrow \mathbb{R}$ does not. This is the case despite the fact that M is locally like \mathbb{R}^n , where differentiability of functions can be defined. If $U \subset M$ is an open set and we choose a homeomorphism $\phi: U \rightarrow \mathbb{R}^n$, it would seem reasonable to define f to be differentiable on U if $f \circ \phi^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable. Unfortunately, if $\psi: V \rightarrow \mathbb{R}^n$ is another homeomorphism, and $U \cap V \neq \emptyset$, then it is not necessarily true that $f \circ \psi^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}$ is also differentiable. Indeed, since

$$f \circ \psi^{-1} = f \circ \phi^{-1} \circ (\phi \circ \psi^{-1}),$$

we can expect $f \circ \psi^{-1}$ to be differentiable for all f which make $f \circ \phi^{-1}$ differentiable only if $\phi \circ \psi^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is differentiable. This is certainly not always



the case; for example, one need merely choose ϕ to be $h \circ \psi$, where $h: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a homeomorphism that is not differentiable.

If we insist on defining differentiable functions on any manifold, there is no way out of this impasse. It is necessary to adorn our manifolds with a little additional structure, the precise nature of which is suggested by the previous discussion.

Among all possible homeomorphisms from $U \subset M$ onto \mathbb{R}^n , we wish to select a certain collection with the property that $\phi \circ \psi^{-1}$ is differentiable whenever ϕ, ψ are in the collection. This is precisely what we shall do, but a few refinements will be introduced along the way.

First of all, we will be interested almost exclusively in functions $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ which are C^∞ (that is, each component function f^i possesses continuous partial derivatives of all orders); sometimes we will use the words “differentiable” or “smooth” to mean C^∞ .

Moreover, instead of considering homeomorphisms from open subsets U of M onto \mathbb{R}^n , it will suffice to consider homeomorphisms $x: U \rightarrow x(U) \subset \mathbb{R}^n$ onto open subsets of \mathbb{R}^n .

The use of the letters x, y , etc., for these homeomorphisms, henceforth adhered to almost religiously, is meant to encourage the casual confusion of a point $p \in M$ with $x(p) \in \mathbb{R}^n$, which has “coordinates” $x^1(p), \dots, x^n(p)$. The only time this notation will be confusing (and it will be) is when we are referring to the manifold \mathbb{R}^n , where it is hard not to lapse back into the practice of denoting points by x and y . We will often mention the pair (x, U) , instead of x alone, just to provide a convenient name for the domain of x .

If U and V are open subsets of M , two homeomorphisms $x: U \rightarrow x(U) \subset \mathbb{R}^n$ and $y: V \rightarrow y(V) \subset \mathbb{R}^n$ are called C^∞ -related if the maps

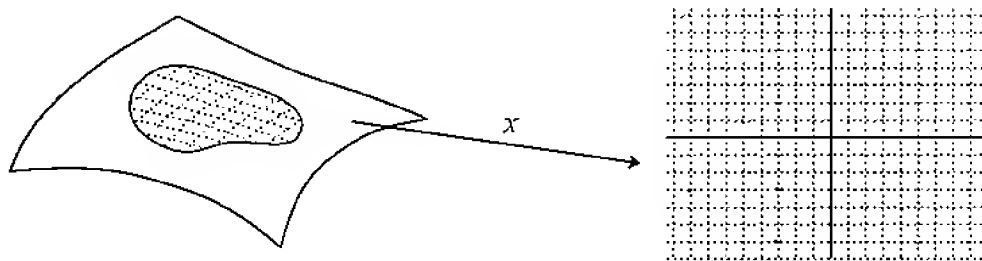
$$\begin{aligned} y \circ x^{-1}: x(U \cap V) &\rightarrow y(U \cap V) \\ x \circ y^{-1}: y(U \cap V) &\rightarrow x(U \cap V) \end{aligned}$$

are C^∞ . This makes sense, since $x(U \cap V)$ and $y(U \cap V)$ are open subsets of \mathbb{R}^n . Also, it makes sense, and is automatically true, if $U \cap V = \emptyset$.

A family of mutually C^∞ -related homeomorphisms whose domains cover M is called an atlas for M . A particular member (x, U) of an atlas \mathcal{A} is called a chart (for the atlas \mathcal{A}), or a coordinate system on U , for the obvious reason that it provides a way of assigning “coordinates” to points on U , namely, the coordinates $x^1(p), \dots, x^n(p)$ to the point $p \in U$.

We can even imagine a mesh of coordinate lines on U , by considering the

inverse images under x of lines in \mathbb{R}^n parallel to one of the axes.



The simplest example of a manifold together with an atlas consists of \mathbb{R}^n with an atlas \mathcal{A} of only one map, the identity $I: \mathbb{R}^n \rightarrow \mathbb{R}^n$. We can easily make the atlas bigger; if U and V are homeomorphic open subsets of \mathbb{R}^n , we can adjoin any homeomorphism $x: U \rightarrow V$ with the property that x and x^{-1} are C^∞ . Indeed, we can adjoin as many such x 's as we like—it is easy to check that they are all C^∞ -related to each other. The advantage of this bigger atlas \mathcal{U} is that the single word “chart”, when applied to this atlas, denotes something which must be described in cumbersome language if one can refer only to \mathcal{A} . Aside from this, \mathcal{U} differs only superficially from \mathcal{A} ; one can easily construct \mathcal{U} from \mathcal{A} (and one would be foolish not to do so once and for all). What has just been said for the atlas $\{I\}$ applies to any atlas:

1. LEMMA. If \mathcal{A} is an atlas of C^∞ -related charts on M , then \mathcal{A} is contained in a unique maximal atlas \mathcal{A}' for M .

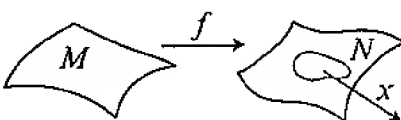
PROOF. Let \mathcal{A}' be the set of all charts y which are C^∞ -related to all charts $x \in \mathcal{A}$. It is easy to check that all charts in \mathcal{A}' are C^∞ -related, so \mathcal{A}' is an atlas, and it is clearly the unique maximal atlas containing \mathcal{A} . ♦

We now define a C^∞ manifold (or differentiable manifold, or smooth manifold) to be a pair (M, \mathcal{A}) , where \mathcal{A} is a maximal atlas for M . Thus, about the simplest example of a C^∞ manifold is $(\mathbb{R}^n, \mathcal{U})$, where \mathcal{U} (the “usual C^∞ -structure for \mathbb{R}^n ”) is the maximal atlas containing $\{I\}$. Another example is $(\mathbb{R}, \mathcal{V})$ where \mathcal{V} contains the homeomorphism $x \mapsto x^3$, whose inverse is *not* C^∞ , together with all charts C^∞ -related to it. Although $(\mathbb{R}, \mathcal{U})$ and $(\mathbb{R}, \mathcal{V})$ are not the same, there is a one-one onto function $f: \mathbb{R} \rightarrow \mathbb{R}$ such that

$$x \in \mathcal{U} \text{ if and only if } x \circ f \in \mathcal{V},$$

namely, the obvious map $f(x) = x^3$. Thus $(\mathbb{R}, \mathcal{U})$ and $(\mathbb{R}, \mathcal{V})$ are the sort of structures one would want to call “isomorphic”. The term actually used is

“diffeomorphic”: two C^∞ manifolds (M, \mathcal{A}) and (N, \mathcal{B}) are **diffeomorphic** if there is a one-one onto function $f: M \rightarrow N$ such that

$$x \in \mathcal{B} \text{ if and only if } x \circ f \in \mathcal{A}.$$


The map f is called a **diffeomorphism**, and f^{-1} is clearly a diffeomorphism also. If we had not required our atlases to be maximal, the definition of diffeomorphism would have had to be more complicated.

Normally, of course, we will suppress mention of the atlas for a differentiable manifold, and speak elliptically of “the differentiable manifold M ”; the atlas for M is sometimes referred to as the *differentiable structure* for M . It will always be understood that \mathbb{R}^n refers to the pair $(\mathbb{R}^n, \mathcal{U})$.

It is easy to see that a diffeomorphism must be continuous. Consequently, its inverse must also be continuous, so that a diffeomorphism is automatically a homeomorphism. This raises the natural question whether, conversely, two homeomorphic manifolds are necessarily diffeomorphic. Later (Problem 9-24) we will be able to prove easily that \mathbb{R} with any atlas is diffeomorphic to $(\mathbb{R}, \mathcal{U})$. A proof of the corresponding assertion for \mathbb{R}^2 is much harder, the proof for \mathbb{R}^3 would certainly be too difficult for inclusion here, and the proof of the essential uniqueness of C^∞ structures on \mathbb{R}^n for $n \geq 5$ requires very difficult techniques from topology.

In the case of spheres, the projections P_1 and P_2 from the points $(0, \dots, 0, 1)$ and $(0, \dots, 0, -1)$ of S^{n-1} are easily seen to be C^∞ -related. They therefore determine an atlas—the “usual C^∞ structure for S^{n-1} ”. This atlas may also be described in terms of the $2n$ homeomorphisms

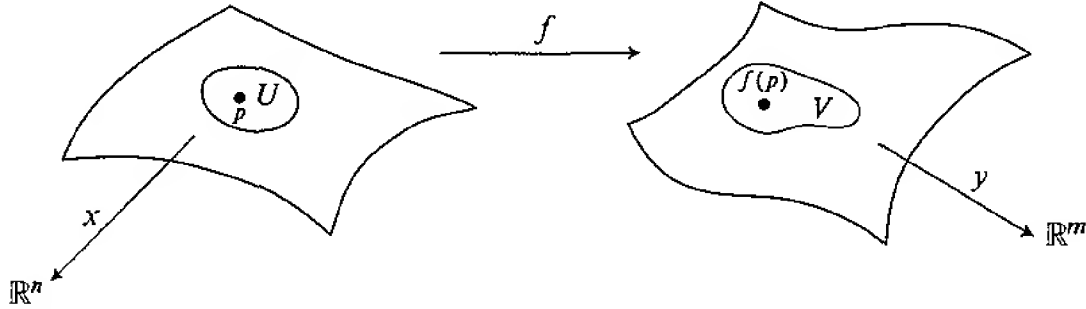
$$\begin{aligned} f_i: S^{n-1} \cap \{x \in \mathbb{R}^n : x^i > 0\} &\rightarrow \mathbb{R}^{n-1} \\ g_i: S^{n-1} \cap \{x \in \mathbb{R}^n : x^i < 0\} &\rightarrow \mathbb{R}^{n-1} \end{aligned}$$

defined by $f_i(x) = g_i(x) = (x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^n)$, which are C^∞ -related to P_1 and P_2 . There are, up to diffeomorphism, unique differentiable structures on S^n for $n \leq 6$. But there are 28 diffeomorphism classes of differentiable structures on S^7 , and over 16 million on S^{31} . However, we shall not come close to proving these assertions, which are part of the field called “differential topology”, rather than differential geometry. (Perhaps most astonishing of all is the quite recent discovery that \mathbb{R}^4 has a differentiable structure that is not diffeomorphic to the usual differentiable structure!)

Other examples of differentiable manifolds will be given soon, but we can already describe a differentiable structure \mathcal{A}' on any open submanifold N of

a differentiable manifold (M, \mathcal{A}) ; the atlas \mathcal{A}' consists of all (x, U) in \mathcal{A} with $U \subset N$.

Just as diffeomorphisms are analogues for C^∞ manifolds of homeomorphisms, there are analogues of continuous maps. A function $f: M \rightarrow N$ is called **differentiable** if for every coordinate system (x, U) for M and (y, V) for N , the map $y \circ f \circ x^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable. More particularly, f



is called **differentiable at $p \in M$** if $y \circ f \circ x^{-1}$ is differentiable at $x(p)$ for coordinate systems (x, U) and (y, V) with $p \in U$ and $f(p) \in V$. If this is true for one pair of coordinate systems, it is easily seen to be true for any other pair. We can thus define differentiability of f on any open subset $M' \subset M$; as one would suspect, this coincides with differentiability of the restricted map $f|_{M'}: M' \rightarrow N$. Clearly, a differentiable map is continuous.

A differentiable function $f: M \rightarrow \mathbb{R}$ refers, of course, to the usual differentiable structure on \mathbb{R} , and hence f is differentiable if and only if $f \circ x^{-1}$ is differentiable for each chart x . It is easy to see that

- (1) a function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable as a map between C^∞ manifolds if and only if it is differentiable in the usual sense;
- (2) a function $f: M \rightarrow \mathbb{R}^m$ is differentiable if and only if each $f^i: M \rightarrow \mathbb{R}^m$ is differentiable;
- (3) a coordinate system (x, U) is a diffeomorphism from U to $x(U)$;
- (4) a function $f: M \rightarrow N$ is differentiable if and only if each $y^i \circ f$ is differentiable for each coordinate system y of N ;
- (5) a differentiable function $f: M \rightarrow N$ is a diffeomorphism if and only if f is one-one onto and $f^{-1}: N \rightarrow M$ is differentiable.

The differentiable structures on many manifolds are designed to make certain functions differentiable. Consider first the product $M_1 \times M_2$ of two differen-

tiabile manifolds M_i , and the two “projections” $\pi_i: M_1 \times M_2 \rightarrow M_i$ defined by $\pi_i(p_1, p_2) = p_i$. It is easy to define a differentiable structure on $M_1 \times M_2$ which makes each π_i differentiable. For each pair (x_i, U_i) of coordinate systems on M_i , we construct the homeomorphism

$$x_1 \times x_2: U_1 \times U_2 \rightarrow \mathbb{R}^{n_1+n_2}$$

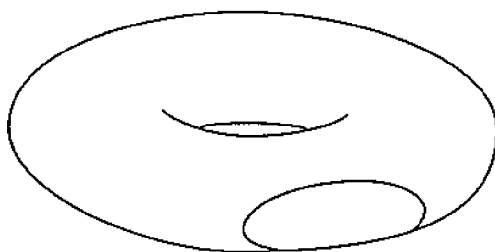
defined by

$$x_1 \times x_2(p_1, p_2) = (x_1(p_1), x_2(p_2)), \quad \text{i.e.,} \quad x_1 \times x_2 = (x_1 \circ \pi_1, x_2 \circ \pi_2).$$

Then we extend this atlas to a maximal one.

Similarly, there is a differentiable structure on \mathbb{P}^n which makes the map $f: S^n \rightarrow \mathbb{P}^n$ (defined by $f(p) = [p] = \{p, -p\}$) differentiable. Consider any coordinate system (x, U) for S^n , where U does *not* contain $-p$ if it contains p , so that $f|U$ is one-one. The map $x \circ (f|U)^{-1}$ is a homeomorphism on $f(U) \subset \mathbb{P}^n$, and any two such are C^∞ -related. The collection of these homeomorphisms can then be extended to a maximal atlas.

To obtain differentiable structures on other surfaces, we first note that a C^∞ manifold-with-boundary can be defined in an obvious way. It is only necessary to know when a map $f: \mathbb{H}^n \rightarrow \mathbb{R}^n$ is to be considered differentiable; we call f **differentiable** when it can be extended to a differentiable function on an open neighborhood of \mathbb{H}^n . A “handle” is then a C^∞ manifold-with-boundary.

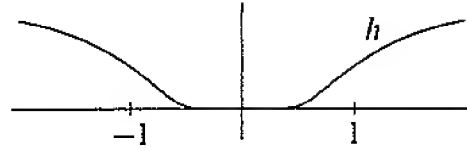


A differentiable structure on the 2-holed torus can be obtained by “matching” the differentiable structure on two handles. The details involved in this process are reserved for Problem 14.

To deal with C^∞ functions effectively, one needs to know that there are lots of them. The existence of C^∞ functions on a manifold depends on the existence of C^∞ functions on \mathbb{R}^n which are 0 outside of a compact set. We briefly recall here the necessary facts about such C^∞ functions (c.f. *Calculus on Manifolds*, pg. 29).

(1) The function $h: \mathbb{R} \rightarrow \mathbb{R}$ defined by

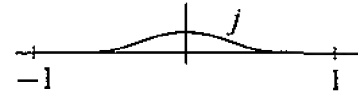
$$h(x) = \begin{cases} e^{-1/x^2} & x \neq 0 \\ 0 & x = 0 \end{cases}$$



is C^∞ , and $h^{(n)}(0) = 0$ for all n .

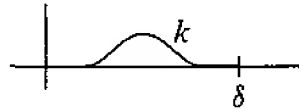
(2) The function $j: \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$j(x) = \begin{cases} e^{-(x-1)^{-2}} \cdot e^{-(x+1)^{-2}} & x \in (-1, 1) \\ 0 & x \notin (-1, 1) \end{cases}$$



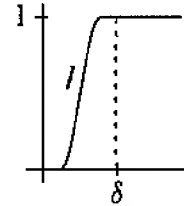
is C^∞ .

Similarly, there is a C^∞ function $k: \mathbb{R} \rightarrow \mathbb{R}$ which is positive on $(0, \delta)$ and 0 elsewhere.



(3) The function $l: \mathbb{R} \rightarrow \mathbb{R}$ defined by

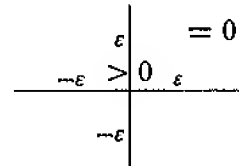
$$l(x) = \left(\int_0^x k \right) / \left(\int_0^\delta k \right)$$



is C^∞ ; it is 0 for $x \leq 0$, increasing on $(0, \delta)$, and 1 for $x \geq \delta$.

(4) The function $g: \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$g(x) = j(x^1/\varepsilon) \cdots j(x^n/\varepsilon)$$

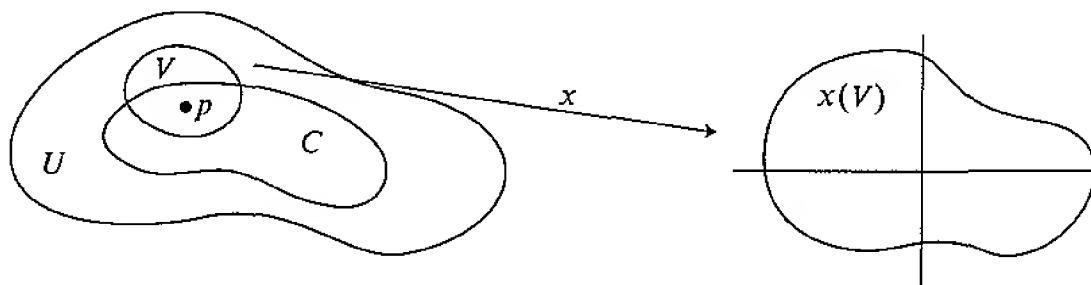


is C^∞ ; it is positive on $(-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon)$ and 0 elsewhere.

On a C^∞ manifold M we can now produce many non-constant C^∞ functions. The closure $\overline{\{x : f(x) \neq 0\}}$ is called the **support** of f , and denoted simply by **support** f (or sometimes $\text{supp } f$).

2. LEMMA. Let $C \subset U \subset M$ with C compact and U open. Then there is a C^∞ function $f: M \rightarrow [0, 1]$ such that $f = 1$ on C and $\text{support } f \subset U$. (Compare *Case 2* of the proof of Theorem 15.)

PROOF. For each $p \in C$, choose a coordinate system (x, V) with $\bar{V} \subset U$ and $x(p) = 0$. Then $x(V) \supset (-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. The function $g \circ x$ (where g is defined in (4)) is C^∞ on V . Clearly it remains C^∞ if we extend



it to be 0 outside of V . Let f_p be the extended function. The function f_p can be constructed for each p , and is positive on a neighborhood of p whose closure is contained in U . Since C is compact, finitely many such neighborhoods cover C , and the sum, $f_{p_1} + \cdots + f_{p_n}$, of the corresponding functions has support $\subset U$. On C it is positive, so on C it is $\geq \delta$ for some $\delta > 0$. Let $f = l \circ (f_{p_1} + \cdots + f_{p_n})$, where l is defined in (3). ♦

By the way, we could have defined C^r manifolds for each $r \geq 1$, not just for “ $r = \infty$ ”. (A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is C^r if it has continuous partial derivatives up to order r). A “ C^0 function” is just a continuous function, so a C^0 manifold is just a manifold in the sense of Chapter 1. We can also define *analytic* manifolds (a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is analytic at $a \in \mathbb{R}^n$ if f can be expressed as a power series in the $(x^i - a^i)$ which converges in some neighborhood of a). The symbol C^ω stands for analytic, and it is convenient to agree that $r < \infty < \omega$ for each integer $r \geq 0$. If $\alpha < \beta$, then the charts of a maximal C^β atlas are all C^α -related, but this atlas can always be extended to a *bigger* atlas of C^α -related charts, as in Lemma 1. Thus, a C^β structure on M can always be extended to a C^α structure in a unique way; the smaller structure is the “stronger” one, the C^0 structure (consisting of all homeomorphisms $x: U \rightarrow \mathbb{R}^n$) being the largest. The converse of this trivial remark is a hard theorem: For $\alpha \geq 1$, every C^α structure contains a C^β structure for each $\beta > \alpha$; it is not unique, of course, but it is unique up to diffeomorphism. This will not be proved here.* In fact, C^α manifolds for $\alpha \neq \infty$ will hardly ever be mentioned again. One remark is in order now; the proof of Lemma 2 produces an appropriate C^α function f on a C^α manifold, for $0 \leq \alpha \leq \infty$. Of course, for $\alpha = \omega$ the proof

* For a proof see Munkres, *Elementary Differential Topology*.

fails completely (and the result is false—an analytic function which is 0 on an open set is 0 everywhere).

With differentiable functions now at our disposal, it is fitting that we begin differentiating them. What we shall define are the partial derivatives of a differentiable function $f: M \rightarrow \mathbb{R}$, with respect to a coordinate system (x, U) . At this point classical notation for partial derivatives is systematically introduced, so it is worth recalling a logical notation for the partial derivatives of a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$. We denote by $D_i f(a)$ the number

$$\lim_{h \rightarrow 0} \frac{f(a^1, \dots, a^i + h, \dots, a^n) - f(a)}{h}.$$

The Chain Rule states that if $g: \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $f: \mathbb{R}^n \rightarrow \mathbb{R}$, then

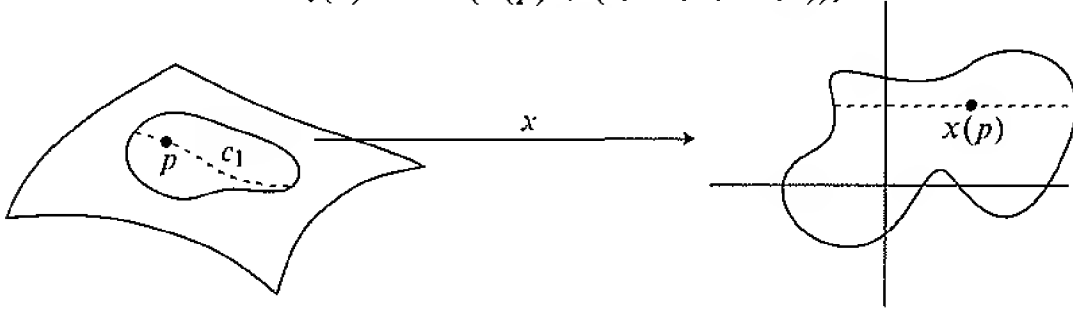
$$D_j(f \circ g)(a) = \sum_{i=1}^n D_i f(g(a)) \cdot D_j g^i(a).$$

Now, for a function $f: M \rightarrow \mathbb{R}$ and a coordinate system (x, U) we define

$$\frac{\partial f}{\partial x^i}(p) = \left. \frac{\partial f}{\partial x^i} \right|_p = D_i(f \circ x^{-1})(x(p)),$$

(or simply $\frac{\partial f}{\partial x^i} = D_i(f \circ x^{-1}) \circ x$, as an equation between functions). If we define the curve $c_i: (-\varepsilon, \varepsilon) \rightarrow M$ by

$$c_i(h) = x^{-1}(x(p) + (0, \dots, h, \dots, 0)),$$



then this partial derivative is just

$$\lim_{h \rightarrow 0} \frac{f(c_i(h)) - f(p)}{h},$$

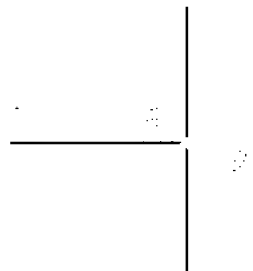
so it measures the rate change of f along the curve c_i ; in fact it is just $(f \circ c_i)'(0)$. Notice that

$$\frac{\partial x^i}{\partial x^j}(p) = \delta_j^i = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

If x happens to be the identity map of \mathbb{R}^n , then $D_i f(p) = \partial f / \partial x^i(p)$, which is the classical symbol for this partial derivative.

Another classical instance of this notation, often not completely clarified, is the use of the symbols $\partial/\partial r$ and $\partial/\partial\theta$ in connection with “polar coordinates”. On the subset A of \mathbb{R}^2 defined by

$$\begin{aligned} A &= \mathbb{R}^2 - \{(x, y) \in \mathbb{R}^2 : y = 0 \text{ and } x \geq 0\} \\ &= \mathbb{R}^2 - L \end{aligned}$$

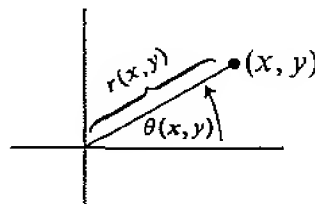


we can introduce a “coordinate system” $P: A \rightarrow \mathbb{R}^2$ by

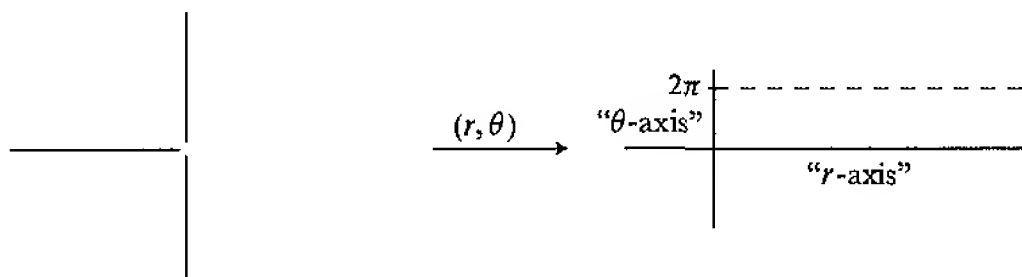
$$P(x, y) = (r(x, y), \theta(x, y)),$$

where $r(x, y) = \sqrt{x^2 + y^2}$ and $\theta(x, y)$ is the unique number in $(0, 2\pi)$ with

$$\begin{aligned} x &= r(x, y) \cos \theta(x, y) \\ y &= r(x, y) \sin \theta(x, y). \end{aligned}$$



This really is a coordinate system on A in our sense, with its image being the set $\{r : r > 0\} \times (0, 2\pi)$. (Of course, the polar coordinate system is often



not restricted to the set A . One can delete any ray other than L if $\theta(x, y)$ is restricted to lie in the appropriate interval $(\theta_0, \theta_0 + 2\pi)$; many results are essentially independent of which line is deleted, and this sometimes justifies the sloppiness involved in the definition of the polar coordinate system.)

We have really defined P as an inverse function, whose inverse P^{-1} is defined simply by

$$P^{-1}(r, \theta) = (r \cos \theta, r \sin \theta).$$

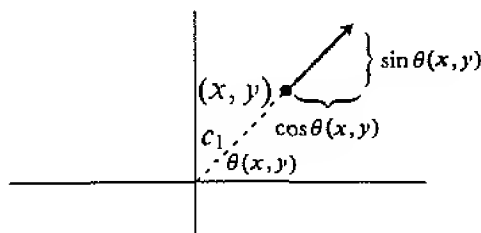
From this formula we can compute $\partial f / \partial r$ explicitly:

$$(f \circ P^{-1})(r, \theta) = f(r \cos \theta, r \sin \theta),$$

so

$$\begin{aligned} \frac{\partial f}{\partial r}(x, y) &= D_1(f \circ P^{-1})(P(x, y)) \\ &= D_1 f(P^{-1}(P(x, y))) \cdot D_1[P^{-1}]^1(P(x, y)) \\ &\quad + D_2 f(P^{-1}(P(x, y))) \cdot D_1[P^{-1}]^2(P(x, y)) \\ &\quad \text{by the Chain Rule} \\ &= D_1 f(x, y) \cdot \cos \theta(x, y) + D_2 f(x, y) \cdot \sin \theta(x, y). \end{aligned}$$

This formula just gives the value of the directional derivative of f at (x, y) , along a unit vector $v = (\cos \theta(x, y), \sin \theta(x, y))$ pointing outwards from the origin to (x, y) . This is to be expected, because c_1 , the inverse image under P

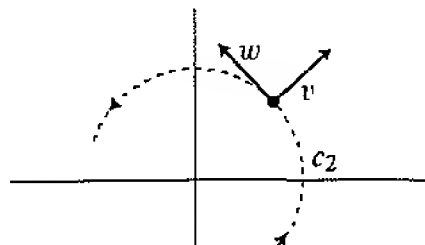


of a curve along the “ r -axis”, is just a line in this direction.

A similar computation gives

$$\frac{\partial f}{\partial \theta}(x, y) = D_1 f(x, y)[-r(x, y) \sin \theta(x, y)] + D_2 f(x, y)[r(x, y) \cos \theta(x, y)].$$

The vector $w = (-\sin \theta(x, y), \cos \theta(x, y))$ is perpendicular to v , and thus the direction, at the point (x, y) , of the curve c_2 which is the inverse image under P of a curve along the “ θ -axis”. The factor $r(x, y)$ appears because this curve



goes around a circle of that radius as θ goes from 0 to 2π , so it is going $r(x, y)$ times as fast as it should go in order to be used to compute the directional derivative of f in the direction w . Note that $\partial f/\partial\theta$ is independent of which line is deleted from the plane in order to define the function θ unambiguously.

Using the notation $\partial f/\partial x$ for $D_1 f$, etc., and suppressing the argument (x, y) everywhere (thus writing an equation about functions), we can write the above equations as

$$\begin{aligned}\frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x} \cos \theta + \frac{\partial f}{\partial y} \sin \theta \\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x} (-r \sin \theta) + \frac{\partial f}{\partial y} r \cos \theta.\end{aligned}$$

In particular, these formulas also tell us what $\partial x/\partial r$ etc., are, where (x, y) denotes the identity coordinate system of \mathbb{R}^2 . We have $\partial x/\partial r = \cos \theta$, etc., so our formulas can be put in the form

$$\begin{aligned}\frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial r} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial r} \\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta}.\end{aligned}$$

In classical notation, the Chain Rule would always be written in this way. It is a pleasure to report that henceforth this may always be done:

3. PROPOSITION. If (x, U) and (y, V) are coordinate systems on M , and $f: M \rightarrow \mathbb{R}$ is differentiable, then on $U \cap V$ we have

$$(I) \quad \frac{\partial f}{\partial y^i} = \sum_{j=1}^n \frac{\partial f}{\partial x^j} \frac{\partial x^j}{\partial y^i}.$$

PROOF. It's the Chain Rule, of course, if you just keep your cool:

$$\begin{aligned}\frac{\partial f}{\partial y^i}(p) &= D_i(f \circ y^{-1})(y(p)) \\ &= D_i([f \circ x^{-1}] \circ [x \circ y^{-1}])(y(p)) \\ &= \sum_{j=1}^n D_j(f \circ x^{-1})([x \circ y^{-1}](y(p))) \cdot D_i[x \circ y^{-1}]^j(y(p))\end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^n D_j(f \circ x^{-1})(x(p)) \cdot D_i[x^j \circ y^{-1}](y(p)) \\
&= \sum_{j=1}^n \frac{\partial f}{\partial x^j}(p) \cdot \frac{\partial x^j}{\partial y^i}(p). \quad \blacklozenge
\end{aligned}$$

At this point we could introduce the “Einstein summation convention”. Notice that the summation in this formula occurs for the index j , which appears both “above” (in $\partial x^j / \partial y^i$) and “below” (in $\partial f / \partial x^j$). There are scads of formulas in which this happens, often with hoards of indices being summed over, and the convention is to omit the \sum sign completely—double indices (which by luck, the nature of things, and felicitous choice of notation, almost always occur above and below) being summed over. I won’t use this notation because whenever I do, I soon forget I’m supposed to be summing, and because by doing things “right”, one can avoid what Élie Cartan has called the “debauch of indices”.

We will often write formula (1) in the form

$$\frac{\partial}{\partial y^i} = \sum_{j=1}^n \frac{\partial x^j}{\partial y^i} \frac{\partial}{\partial x^j};$$

here $\partial / \partial y^i$ is considered as an operator taking the function f to $\partial f / \partial y^i$. The operator taking f to $\partial f / \partial y^i(p)$ is denoted by

$$\left. \frac{\partial}{\partial y^i} \right|_p; \quad \text{thus} \quad \left. \frac{\partial}{\partial y^i} \right|_p = \sum_{j=1}^n \left. \frac{\partial x^j}{\partial y^i}(p) \frac{\partial}{\partial x^j} \right|_p.$$

For later use we record a property of $\ell = \partial / \partial x^i|_p$: it is a “point-derivation”.

4. PROPOSITION. For any differentiable $f, g: M \rightarrow \mathbb{R}$, and any coordinate system (x, U) with $p \in U$, the operator $\ell = \partial / \partial x^i|_p$ satisfies

$$\ell(fg) = f(p)\ell(g) + \ell(f)g(p).$$

PROOF. Left to the reader. \blacklozenge

If (x, U) and (x', U') are two coordinate systems on M , the $n \times n$ matrix

$$\left(\frac{\partial x'^i}{\partial x^j}(p) \right)$$

is just the Jacobian matrix of $x' \circ x^{-1}$ at $x(p)$. It is non-singular; in fact, its inverse is clearly

$$\left(\frac{\partial x^i}{\partial x'^j}(p) \right).$$

Now if $f: M^n \rightarrow N^m$ is C^∞ and (y, V) is a coordinate system around $f(p)$, the rank of the $m \times n$ matrix

$$\left(\frac{\partial(y^i \circ f)}{\partial x^j}(p) \right)$$

clearly does not depend on the coordinate system (x, U) or (y, V) . It is called the **rank of f at p** . The point p is called a **critical point** of f if the rank of f at p is $< m$ (the dimension of the image N); if p is not a critical point of f , it is called a **regular point** of f . If p is a critical point of f , the value $f(p)$ is called a **critical value** of f . Other points in N are **regular values**; thus $q \in N$ is a regular value if and only if p is a regular point of f for every $p \in f^{-1}(q)$. This is true, in particular, if $q \notin f(M)$ —a non-value of f is still a “regular value”.

If $f: \mathbb{R} \rightarrow \mathbb{R}$, then x is a critical point of f if and only if $f'(x) = 0$. It is possible for all points of the interval $[a, b]$ to be critical points, although this can happen only if f is constant on $[a, b]$. If $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ has all points as critical values, then $D_1 f = D_2 f = 0$ everywhere, so f is again constant. On the other hand, a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ may have all points as critical points without being constant, for example, $f(x, y) = x$. In this case, however, the image $f(\mathbb{R}^2) = \mathbb{R} \times \{0\} \subset \mathbb{R}^2$ is still a “small” subset of \mathbb{R}^2 . The most important theorem about critical points generalizes this fact. To state it, we will need some terminology.

Recall that a set $A \subset \mathbb{R}^n$ has “measure zero” if for every $\varepsilon > 0$ there is a sequence B_1, B_2, B_3, \dots of (closed or open) rectangles with

$$A \subset \bigcup_{n=1}^{\infty} B_n$$

and

$$\sum_{n=1}^{\infty} v(B_n) < \varepsilon,$$

where $v(B_n)$ is the volume of B_n . We want to define the same concept for a subset of a manifold. To do this we need a lemma, which in turn depends on a lemma from *Calculus on Manifolds*, which we merely state.

5. LEMMA. Let $A \subset \mathbb{R}^n$ be a rectangle and let $f: A \rightarrow \mathbb{R}^n$ be a function such that $|D_j f^i| \leq K$ on A for $i, j = 1, \dots, n$. Then

$$|f(x) - f(y)| \leq n^2 K |x - y|$$

for all $x, y \in A$.

6. LEMMA. If $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is C^1 and $A \subset \mathbb{R}^n$ has measure 0, then $f(A)$ has measure 0.

PROOF. We can assume that A is contained in a compact set C (since \mathbb{R}^n is a countable union of compact sets). Lemma 5 implies that there is some K such that

$$|f(x) - f(y)| \leq n^2 K |x - y|$$

for all $x, y \in C$. Thus f takes rectangles of diameter d into sets of diameter $\leq n^2 K d$. This clearly implies that $f(A)$ has measure 0 if A does. ♦

A subset A of a C^∞ n -manifold M has **measure zero** if there is a sequence of charts (x_i, U_i) , with $A \subset \bigcup_i U_i$, such that each set $x_i(A \cap U_i) \subset \mathbb{R}^n$ has measure 0. Using Lemma 6, it is easy to see that if $A \subset M$ has measure 0, then $x(A \cap U) \subset \mathbb{R}^n$ has measure 0 for any coordinate system (x, U) . Conversely, if this condition is satisfied and M is connected, or has only countably many components, then it follows easily from Theorem 1-2 that A has measure 0. (But if M is the disjoint union of uncountably many copies of \mathbb{R} , and A consists of one point from each component, then A does not have measure 0). Lemma 6 thus implies another result:

7. COROLLARY. If $f: M \rightarrow N$ is a C^1 function between two n -manifolds and $A \subset M$ has measure 0, then $f(A) \subset N$ has measure 0.

PROOF. There is a sequence of charts (x_i, U_i) with $A \subset \bigcup_i U_i$ and each set $x_i(A \cap U_i)$ of measure 0. If (y, V) is a chart on N , then $f(A) \cap V = \bigcup_i f(A \cap U_i) \cap V$. Each set

$$y(f(A \cap U_i) \cap V) = y \circ f \circ x^{-1}(x(A \cap U_i))$$

has measure 0, by Lemma 6. Thus $y(f(A) \cap V)$ has measure 0. Since $f(\bigcup_i U_i)$ is contained in the union of at most countably many components of N , it follows that $f(A)$ has measure 0. ♦

8. THEOREM (SARD'S THEOREM). If $f: M \rightarrow N$ is a C^1 map between n -manifolds, and M has at most countably many components, then the critical values of f form a set of measure 0 in N .

PROOF. It clearly suffices to consider the case where M and N are \mathbb{R}^n . But this case is just Theorem 3.14 of *Calculus on Manifolds*. ♦

The stronger version of Sard's Theorem, which we will never use (except once, in Problem 8-24), states* that the critical values of a C^k map $f: M^n \rightarrow N^m$ are a set of measure 0 if $k \geq 1 + \max(n - m, 0)$. Theorem 8 is the easy case, and the case $m > n$ is the trivial case (Problem 20). Although Theorem 8 will be very important later on, for the present we are more interested in knowing what the image of $f: M \rightarrow N$ looks like locally, in terms of the rank k of f at $p \in M$. More exact information can be given when f actually has rank k in a neighborhood of p . It should be noted that f must have rank $\geq k$ in some neighborhood of p , because some $k \times k$ submatrix of $(\partial(y^i \circ f)/\partial x^i)$ has non-zero determinant at p , and hence in a neighborhood of p .

9. THEOREM. (1) If $f: M^n \rightarrow N^m$ has rank k at p , then there is some coordinate system (x, U) around p and some coordinate system (y, V) around $f(p)$ with $y \circ f \circ x^{-1}$ in the form

$$y \circ f \circ x^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^k, \psi^{k+1}(a), \dots, \psi^m(a)).$$

Moreover, given any coordinate system y , the appropriate coordinate system on N can be obtained merely by permuting the component functions of y .

(2) If f has rank k in a neighborhood of p , then there are coordinate systems (x, U) and (y, V) such that

$$y \circ f \circ x^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^k, 0, \dots, 0).$$

Remark: The special case $M = \mathbb{R}^n$, $N = \mathbb{R}^m$ is equivalent to the general theorem, which gives only local results. If y is the identity of \mathbb{R}^m , part (1) says that by first performing a diffeomorphism on \mathbb{R}^n , and then permuting the coordinates in \mathbb{R}^m , we can insure that f keeps the first k components of a point fixed. These diffeomorphisms on \mathbb{R}^n and \mathbb{R}^m are clearly necessary, since f may not even be one-one on $\mathbb{R}^k \times \{0\} \subset \mathbb{R}^n$, and its image could, for example, contain only points with first coordinate 0.

*For a proof, see Milnor, *Topology From the Differentiable Viewpoint* or Sternberg, *Lectures on Differential Geometry*.

In part (2) we must clearly allow more leeway in the choice of y , since $f(\mathbb{R}^n)$ may not be contained in any k -dimensional subspace of \mathbb{R}^n .

PROOF. (1) Choose some coordinate system u around p . By a permutation of the coordinate functions u^i and y^i we can arrange that

$$(1) \quad \det \left(\frac{\partial(y^\alpha \circ f)}{\partial u^\beta}(p) \right) \neq 0 \quad \alpha, \beta = 1, \dots, k.$$

Define

$$\begin{aligned} x^\alpha &= y^\alpha \circ f & \alpha &= 1, \dots, k \\ x^r &= u^r & r &= k+1, \dots, n. \end{aligned}$$

Condition (1) implies that

$$(2) \quad \det \left(\frac{\partial x^i}{\partial u^j}(p) \right) = \det \left(\begin{array}{c|c} \frac{\partial(y^\alpha \circ f)}{\partial u^\beta} & \begin{matrix} \times \\ \times \\ \times \end{matrix} \\ \hline 0 & \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \end{array} \right) \neq 0.$$

This shows that $x = (x \circ u^{-1}) \circ u$ is a coordinate system in some neighborhood of p , since (2) and the Inverse Function Theorem show that $x \circ u^{-1}$ is a diffeomorphism in a neighborhood of $u(p)$. Now

$$\begin{aligned} q = x^{-1}(a^1, \dots, a^n) & \text{ means } x(q) = (a^1, \dots, a^n), \\ \text{hence } x^i(q) &= a^i, \\ \text{hence } \begin{cases} y^\alpha \circ f(q) = a^\alpha & \alpha = 1, \dots, k \\ u^r(q) = a^r & r = k+1, \dots, n, \end{cases} \end{aligned}$$

so

$$\begin{aligned} y \circ f \circ x^{-1}(a^1, \dots, a^n) &= y \circ f(q) \quad \text{for } q = x^{-1}(a^1, \dots, a^n) \\ &= (a^1, \dots, a^k, \underline{\quad}). \end{aligned}$$

(2) Choose coordinate systems x and v so that $v \circ f \circ x^{-1}$ has the form in (1). Since $\text{rank } f = k$ in a neighborhood of p , the lower square in the matrix

$$\left(\frac{\partial(v^i \circ f)}{\partial x^j} \right) = \left(\begin{array}{c|c} \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} & 0 \\ \hline \begin{matrix} \times \\ \times \\ \times \end{matrix} & \begin{matrix} D_{k+1}\psi^{k+1} & & \\ & \ddots & \\ & & D_m\psi^m \end{matrix} \end{array} \right)$$

must vanish in a neighborhood of p . Thus we can write

$$\psi^r(a) = \tilde{\psi}^r(a^1, \dots, a^k) \quad r = k+1, \dots, m.$$

Define

$$\begin{aligned} y^\alpha &= v^\alpha \\ y^r &= v^r - \tilde{\psi}^r \circ (v^1, \dots, v^k). \end{aligned}$$

Since

$$\begin{aligned} (3) \quad y \circ v^{-1}(b^1, \dots, b^m) &= y(q) \quad \text{for } v(q) = (b^1, \dots, b^m) \\ &= (b^1, \dots, b^k, b^{k+1} - \tilde{\psi}^{k+1}(b^1, \dots, b^k), \dots, b^m - \tilde{\psi}^m(b^1, \dots, b^k)), \end{aligned}$$

the Jacobian matrix

$$\left(\frac{\partial y^i}{\partial v^j} \right) = \begin{pmatrix} \boxed{\begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix}} & 0 \\ \text{X} & \boxed{\begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix}} \end{pmatrix}$$

has non-zero determinant, so y is a coordinate system in a neighborhood of $f(p)$. Moreover,

$$\begin{aligned} y \circ f \circ x^{-1}(a^1, \dots, a^n) &= y \circ v^{-1} \circ v \circ f \circ x^{-1}(a^1, \dots, a^n) \\ &= y \circ v^{-1}(a^1, \dots, a^k, \psi^{k+1}(a), \dots, \psi^m(a)) \\ &= (a^1, \dots, a^k, \psi^{k+1}(a) - \tilde{\psi}^{k+1}(a^1, \dots, a^k), \dots, \psi^m(a) - \tilde{\psi}^m(a^1, \dots, a^k)) \\ &\quad \text{by (3)} \\ &= (a^1, \dots, a^k, 0, \dots, 0). \quad \spadesuit \end{aligned}$$

Theorem 9 acquires a special form when the rank of f is n or m :

10. THEOREM. (1) If $m \leq n$ and $f: M^n \rightarrow N^m$ has rank m at p , then for any coordinate system (y, V) around $f(p)$, there is some coordinate system (x, U) around p with

$$y \circ f \circ x^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^m).$$

(2) If $n \leq m$ and $f: M^n \rightarrow N^m$ has rank n at p , then for any coordinate system (x, U) around p , there is a coordinate system (y, V) around $f(p)$ with

$$y \circ f \circ x^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^n, 0, \dots, 0).$$

PROOF. (1) This is practically a special case of (1) in Theorem 9; it is only necessary to observe that when $k = m$, it is clearly unnecessary, in the proof of this case, to permute the y^i in order to arrange that

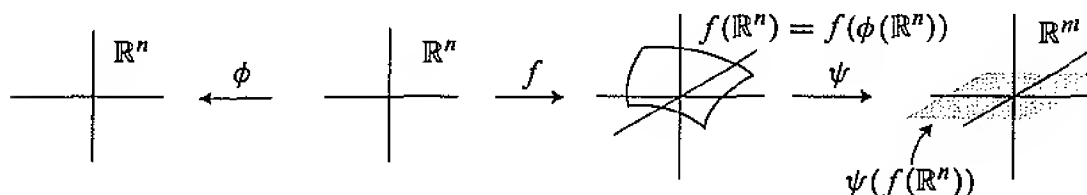
$$\det \left(\frac{\partial (y^\alpha \circ f)}{\partial u^\beta} (p) \right) \neq 0 \quad \alpha, \beta = 1, \dots, m;$$

only the u^i need be permuted.

(2) Since the rank of f at any point must be $\leq n$, the rank of f equals n in some neighborhood of p . It is convenient to think of the case $M = \mathbb{R}^n$ and $N = \mathbb{R}^m$ and produce the coordinate system y for \mathbb{R}^m when we are given the identity coordinate system for \mathbb{R}^n . Part (2) of Theorem 9 yields coordinate systems ϕ for \mathbb{R}^n and ψ for \mathbb{R}^m such that

$$\psi \circ f \circ \phi^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^n, 0, \dots, 0).$$

Even if we do not perform ϕ^{-1} first, the map f still takes \mathbb{R}^n into the subset



$f(\mathbb{R}^n)$ which ψ takes to $\mathbb{R}^n \times \{0\} \subset \mathbb{R}^m$ —the points of \mathbb{R}^n just get moved to the wrong place in $\mathbb{R}^n \times \{0\}$. This can be corrected by another map on \mathbb{R}^m . Define λ by

$$\lambda(b^1, \dots, b^m) = (\phi^{-1}(b^1, \dots, b^n), b^{n+1}, \dots, b^m).$$

Then

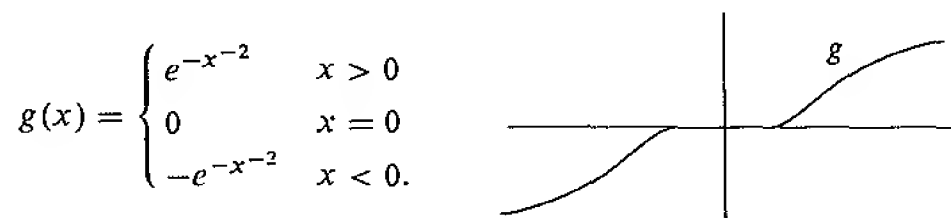
$$\begin{aligned} \lambda \circ \psi \circ f(a^1, \dots, a^n) &= \lambda \circ \psi \circ f \circ \phi^{-1}(b^1, \dots, b^n) \\ &\quad \text{for } (b^1, \dots, b^n) = \phi(a) \\ &= \lambda(b^1, \dots, b^n, 0, \dots, 0) \\ &= (\phi^{-1}(b^1, \dots, b^n), 0, \dots, 0) \\ &= (a^1, \dots, a^n, 0, \dots, 0), \end{aligned}$$

so $\lambda \circ \psi$ is the desired y . If we are given a coordinate system x on \mathbb{R}^n other than the identity, we just define

$$\lambda(b^1, \dots, b^m) = (x(\phi^{-1}(b^1, \dots, b^n)), b^{n+1}, \dots, b^m);$$

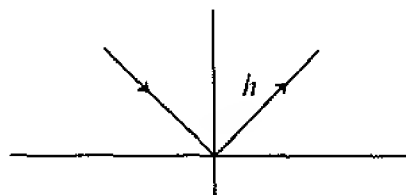
it is easily checked that $y = \lambda \circ \psi$ is now the desired y . ♦

Although p is a regular point of f in case (1) of Theorem 10 and a critical point in case (2) (if $n < m$), it is case (2) which most interests us. A differentiable function $f: M^n \rightarrow N^m$ is called an **immersion** if the rank of f is n , the dimension of the domain M , at all points of M . Of course, it is necessary that $m \geq n$, and it is clear from Theorem 10(2) that an immersion is locally one-one (so it is a topological immersion, as defined in Chapter 1). On the other hand, a differentiable map f need not be an immersion even if it is globally one-one. The simplest example is the function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x^3$, with $f'(0) = 0$. Another example is



A more illuminating example is the function $h: \mathbb{R} \rightarrow \mathbb{R}^2$ defined by

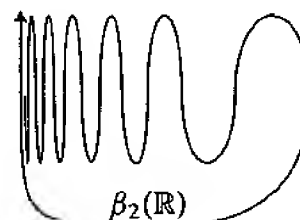
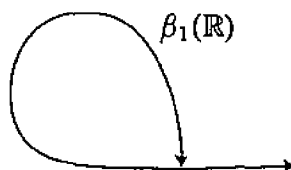
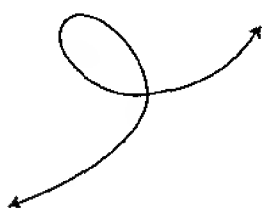
$$h(x) = (g(x), |g(x)|);$$



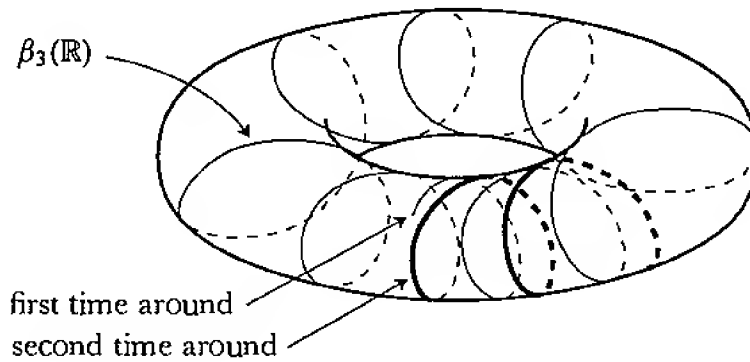
although its image is the graph of a non-differentiable function, the curve itself manages to be differentiable by slowing down to velocity 0 at the point $(0,0)$. One can easily define a similar curve whose image looks like the picture below.



Three immersions of \mathbb{R} in \mathbb{R}^2 are shown below. Although the second and third immersions β_1 and β_2 are one-one, their images are not homeomorphic



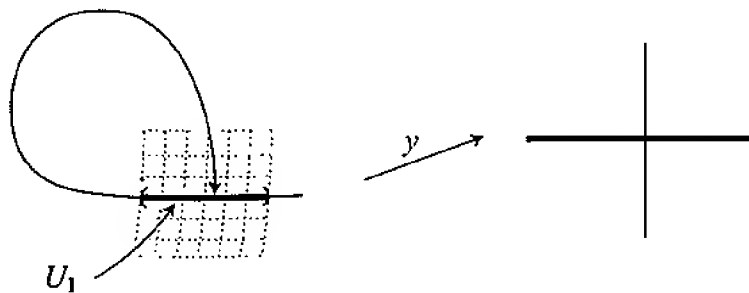
to \mathbb{R} . Of course, even if the one-one immersion $f: P \rightarrow M$ is not a homeomorphism onto its image, there is certainly some metric and some differentiable structure on $f(P)$ which makes the inclusion map $i: f(P) \rightarrow M$ an immersion. In general, a subset $M_1 \subset M$, with a differentiable structure (not necessarily compatible with the metric M_1 inherits as a subset of M), is called an **immersed submanifold** of M if the inclusion map $i: M_1 \rightarrow M$ is an immersion. The following picture, indicating the image of an immersion $\beta_3: \mathbb{R} \rightarrow S^1 \times S^1$,



shows that M_1 may even be a dense subset of M .

Despite these complications, if M_1 is a k -dimensional immersed submanifold of M^n and U_1 is a neighborhood in M_1 of a point $p \in M_1$, then there is a coordinate system (y, V) of M around p , such that

$$U_1 \cap V = \{q \in M : y^{k+1}(q) = \dots = y^n(q) = 0\};$$



this is an immediate consequence of Theorem 10(2), with $f = i$. Thus, if $g: M_1 \rightarrow N$ is C^∞ (considered as a function on the manifold M_1) in a neighborhood of a point $p \in M_1$, then there is a C^∞ function \bar{g} on a neighborhood $V \subset M$ of p such that $g = \bar{g} \circ i$ on $V \cap M_1$ —we can define

$$\bar{g}(q) = g(q'), \quad \text{where} \quad \begin{cases} y^\alpha(q') = y^\alpha(q) & \alpha = 1, \dots, k \\ y^r(q') = 0 & r = k+1, \dots, n. \end{cases}$$

On the other hand, even if g is C^∞ on all of M_1 we may not be able to define \tilde{g} on M . For example, this cannot be done if g is one of the functions $\beta_i^{-1}: \beta_i(M) \rightarrow \mathbb{R}$.

One other complication arises with immersed submanifolds. If $M_1 \subset M$ is an immersed submanifold, and $f: P \rightarrow M$ is a C^∞ function with $f(P) \subset M_1$, it is not necessarily true that f is C^∞ when considered as a map into M_1 , with its C^∞ structure. The following figure shows that f might not even be continuous



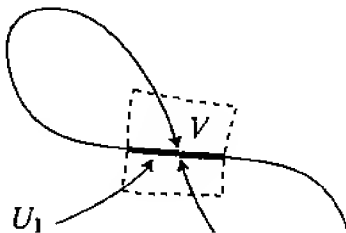
as a map into M_1 . Actually, this is the only thing that can go wrong:

11. PROPOSITION. If $M_1 \subset M$ is an immersed manifold, $f: P \rightarrow M$ is a C^∞ function with $f(P) \subset M_1$, and f is continuous considered as a map into M_1 , then f is also C^∞ considered as a map into M_1 .

PROOF. Let $i: M_1 \rightarrow M$ be the inclusion map. We want to show that $i^{-1} \circ f$ is C^∞ if it is continuous. Given $p \in P$, choose a coordinate system (y, V) for M around $f(p)$ such that

$$U_1 = \{q \in V : y^{k+1}(q) = \cdots = y^n(q) = 0\}$$

is a neighborhood of $f(p)$ in M_1 and $(y^1|_{U_1}, \dots, y^k|_{U_1})$ is a coordinate system of M_1 on U_1 .

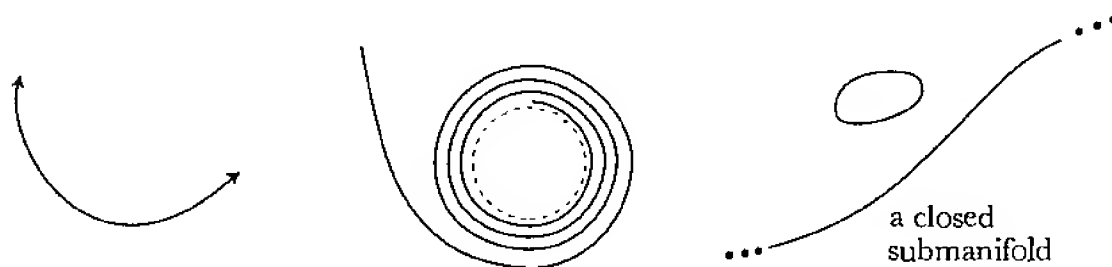


By assumption, $i^{-1} \circ f$ is continuous, so

$$f^{-1} \circ i(\text{open set}) \text{ is an open set.}$$

Since U_1 is open in M_1 , this means that $f^{-1}(U_1) \subset P$ is open. Thus f takes some neighborhood of $p \in P$ into U_1 . Since all $y^j \circ f$ are C^∞ , and y^1, \dots, y^k are a coordinate system on U_1 , the function f is C^∞ considered as a map into M_1 . ♦

Most of these difficulties disappear when we consider one-one immersions $f: P \rightarrow M$ which are homeomorphisms onto their image. Such an immersion is called an *imbedding* ("embedding" for the English). An immersed submanifold $M_1 \subset M$ is called simply a (C^∞) *submanifold* of M if the inclusion map $i: M_1 \rightarrow M$ is an imbedding; it is called a *closed submanifold* of M if M_1 is also a closed subset of M .



There is one way of getting submanifolds which is very important, and gives the sphere $S^{n-1} \subset \mathbb{R}^n - \{0\} \subset \mathbb{R}^n$, defined as $\{x : |x|^2 = 1\}$, as a special case.

12. PROPOSITION. If $f: M^n \rightarrow N$ has constant rank k on a neighborhood of $f^{-1}(y)$, then $f^{-1}(y)$ is a closed submanifold of M of dimension $n - k$ (or is empty). In particular, if y is a regular value of $f: M^n \rightarrow N^m$, then $f^{-1}(y)$ is an $(n - m)$ -dimensional submanifold of M (or is empty).

PROOF. Left to the reader. ♦

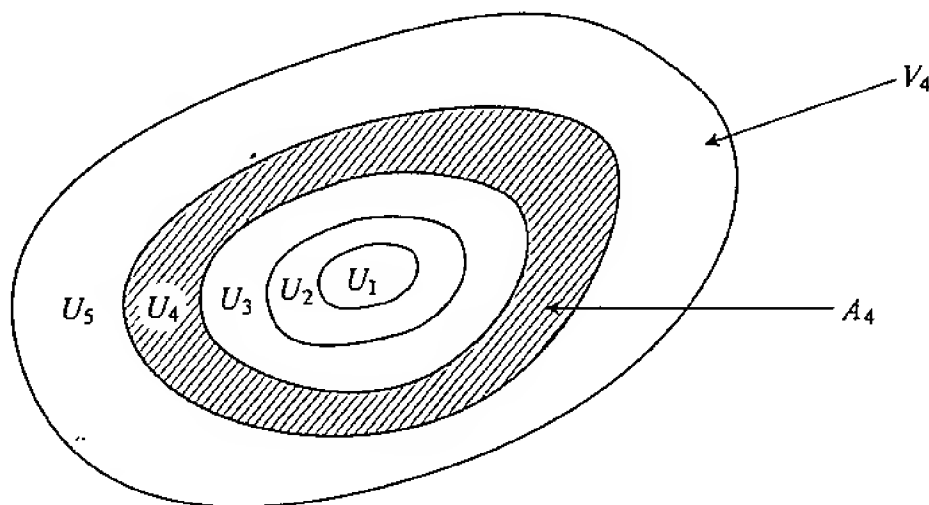
It is to be hoped that however abstract the notion of C^∞ manifolds may appear, submanifolds of \mathbb{R}^N will seem like fairly concrete objects. Now it turns out that *every* (connected) C^∞ manifold can be imbedded in some \mathbb{R}^N , so that manifolds can be pictured as subsets of Euclidean space (though this picture is not always the most useful one). We will prove this fact only for compact manifolds, but we first develop some of the machinery which would be used in

the general case, since we will need it later on anyway. Unfortunately, there are many definitions and theorems involved.

If \mathcal{O} is a cover of a space M , a cover \mathcal{O}' of M is a **refinement** of \mathcal{O} (or “refines \mathcal{O} ”) if for every U in \mathcal{O}' there is some V in \mathcal{O} with $U \subset V$ (the sets of \mathcal{O}' are “smaller” than those of \mathcal{O})—a subcover is a very special case of a refining cover. A cover \mathcal{O} is called **locally finite** if every $p \in M$ has a neighborhood W which intersects only finitely many sets in \mathcal{O} .

13. THEOREM. If \mathcal{O} is an open cover of a manifold M , then there is an open cover \mathcal{O}' of M which is locally finite and which refines \mathcal{O} . Moreover, we can choose all members of \mathcal{O}' to be open sets diffeomorphic to \mathbb{R}^n .

PROOF. We can obviously assume that M is connected. By Theorem 1.2, there are compact sets C_1, C_2, C_3, \dots with $M = C_1 \cup C_2 \cup C_3 \cup \dots$. Clearly C_1 has an open neighborhood U_1 with compact closure. Then $\overline{U_1} \cup C_2$ has an open neighborhood U_2 with compact closure. Continuing in this way, we obtain open sets U_i , with $\overline{U_i}$ compact and $\overline{U_i} \subset U_{i+1}$, whose union contains all C_i , and hence is M . Let $U_{-1} = U_0 = \emptyset$.



Now M is the union for $i > 1$ of the “annular” regions $A_i = \overline{U_i} - U_{i-1}$. Since each A_i is compact, we can obviously cover A_i by a finite number of open sets, each contained in some member of \mathcal{O} , and each contained in $V_i = U_{i+1} - U_{i-2}$. We can also choose these open sets to be diffeomorphic to \mathbb{R}^n . In this way we obtain a cover \mathcal{O}' which refines \mathcal{O} and which is locally finite, since a point in U_i is not in V_j for $j \geq 2 + i$. ♦

Notice that if \mathcal{O} is an open locally finite cover of a space M and $C \subset M$ is compact, then C intersects only finitely many members of \mathcal{O} . This shows that an open locally finite cover of a connected manifold must be countable (like the cover constructed in the proof of Theorem 13).

14. THEOREM (THE SHRINKING LEMMA). Let \mathcal{O} be an open locally finite cover of a manifold M . Then it is possible to choose, for each U in \mathcal{O} , an open set U' with $\overline{U'} \subset U$ in such a way that the collection of all U' is also an open cover of M .

PROOF. We can clearly assume that M is connected. Let $\mathcal{O} = \{U_1, U_2, U_3, \dots\}$. Then

$$C_1 = U_1 - (U_2 \cup U_3 \cup \dots)$$

is a closed set contained in U_1 , and $M = C_1 \cup U_2 \cup U_3 \cup \dots$. Let U'_1 be an open set with $C_1 \subset U'_1 \subset \overline{U'_1} \subset U_1$. Now

$$C_2 = U_2 - (U'_1 \cup U_3 \cup \dots)$$

is a closed set contained in U_2 , and $M = U'_1 \cup C_2 \cup U_3 \cup \dots$. Let U'_2 be an open set with $C_2 \subset U'_2 \subset \overline{U'_2} \subset U_2$. Continue in this way.

For any $p \in M$ there is a largest n with $p \in U_n$, because \mathcal{O} is locally finite. Now

$$p \in U'_1 \cup U'_2 \cup \dots \cup U'_n \cup (U_{n+1} \cup U_{n+2} \cup \dots);$$

it follows that

$$p \in U'_1 \cup U'_2 \cup \dots,$$

since replacing U_{n+i} by U'_{n+i} cannot possibly eliminate p . ♦

15. THEOREM. Let \mathcal{O} be an open locally finite cover of a manifold M . Then there is a collection of C^∞ functions $\phi_U: M \rightarrow [0, 1]$, one for each U in \mathcal{O} , such that

- (1) support $\phi_U \subset U$ for each U ,
- (2) $\sum_U \phi_U(p) = 1$ for all $p \in M$ (this sum is really a finite sum in some neighborhood of p , by (1)).

PROOF. Case 1. Each U in \mathcal{O} has compact closure. Choose the U' as in Theorem 14. Apply Lemma 2 to $\overline{U'} \subset U \subset M$ to obtain a C^∞ function $\psi_U: M \rightarrow [0, 1]$ which is 1 on $\overline{U'}$ and has support $\subset U$. Since the U' cover M , clearly

$$\sum_{U \in \mathcal{O}} \psi_U > 0 \quad \text{everywhere.}$$

Define

$$\phi_U = \frac{\psi_U}{\sum_{U \in \mathcal{O}} \psi_U}.$$

Case 2. General case. This case can be proved in the same way, provided that Lemma 2 is true for $C \subset U \subset M$ with C closed (but not necessarily compact) and U open. But this is a consequence of *Case 1*:

For each $p \in C$ choose an open set $U_p \subset U$ with compact closure. Cover $M - C$ with open sets V_α having compact closure and contained in $M - C$. The open cover $\{U_p, V_\alpha\}$ has an open locally finite refinement \mathcal{O} to which *Case 1* applies. Let

$$f = \sum_{U \in \mathcal{O}'} \phi_U, \quad \text{where } \mathcal{O}' = \{U \in \mathcal{O} : U \subset U_p \text{ for some } p\}.$$

This sum is C^∞ , since it is a finite sum in a neighborhood of each point. Since $\sum_U \phi_U(p) = 1$ for all p , and $\phi_U(p) = 0$ when $U \subset V_\alpha$, clearly $f(p) = 1$ for all $p \in C$. Using the fact that \mathcal{O} is locally finite, it is easy to see that support $f \subset U$. ♦

16. COROLLARY. If \mathcal{O} is any open cover of a manifold M , then there is a collection of C^∞ functions $\phi_i: M \rightarrow [0, 1]$ such that

- (1) the collection of sets $\{p: \phi_i(p) \neq 0\}$ is locally finite,
- (2) $\sum_i \phi_i(p) = 1$ for all $p \in M$,
- (3) for each i there is a $U \in \mathcal{O}$ such that support $\phi_i \subset U$.

(A collection $\{\phi_i: M \rightarrow [0, 1]\}$ satisfying (1) and (2) is called a **partition of unity**; if it satisfies (3), it is called **subordinate to \mathcal{O}** .)

It is now fairly easy to prove the last theorem of this chapter.

17. THEOREM. If M^n is a compact C^∞ manifold, then there is an imbedding $f: M \rightarrow \mathbb{R}^N$ for some N .

PROOF. There are a finite number of coordinate systems $(x_1, U_1), \dots, (x_k, U_k)$ with $M = U_1 \cup \dots \cup U_k$. Choose U'_i as in Theorem 14, and functions $\psi_i: M \rightarrow [0, 1]$ which are 1 on $\overline{U'_i}$ and have support $\subset U_i$. Define $f: M \rightarrow \mathbb{R}^N$, where $N = nk + k$, by

$$f = (\psi_1 \cdot x_1, \dots, \psi_k \cdot x_k, \psi_1, \dots, \psi_k).$$

This is an immersion, because any point p is in U'_i for some i , and on U'_i , where $\psi_i = 1$, the $N \times n$ Jacobian matrix

$$\left(\frac{\partial f^\alpha}{\partial x_i^\beta} \right) \text{ contains the } n \times n \text{ matrix } \left(\frac{\partial x_i^\alpha}{\partial x_i^\beta} \right) = I.$$

It is also one-one. For suppose that $f(p) = f(q)$. There is some i such that $p \in U'_i$. Then $\psi_i(p) = 1$, so also $\psi_i(q) = 1$. This shows that we must have $q \in U_i$. Moreover,

$$\psi_i \cdot x_i(p) = \psi_i \cdot x_i(q),$$

so $p = q$, since x_i is one-one on U_i . ♦

Problem 3-33 shows that, in fact, we can always choose $N = 2n + 1$.

PROBLEMS

1. (a) Show that being C^∞ -related is *not* an equivalence relation.
 (b) In the proof of Lemma 1, show that all charts in \mathcal{A}' are C^∞ -related, as claimed.
2. (a) If M is a metric space together with a collection of homeomorphisms $x: U \rightarrow \mathbb{R}^n$ whose domains cover M and which are C^∞ -related, show that the n at each point is unique *without* using Invariance of Domain.
 (b) Show similarly that ∂M is well-defined for a C^∞ manifold-with-boundary M .
3. (a) All C^∞ functions are continuous, and the composition of C^∞ functions is C^∞ .
 (b) A function $f: M \rightarrow N$ is C^∞ if and only if $g \circ f$ is C^∞ for every C^∞ function $g: N \rightarrow \mathbb{R}$.
4. How many distinct C^∞ structures are there on \mathbb{R} ? (There is only one up to diffeomorphism; that is not the question being asked.)

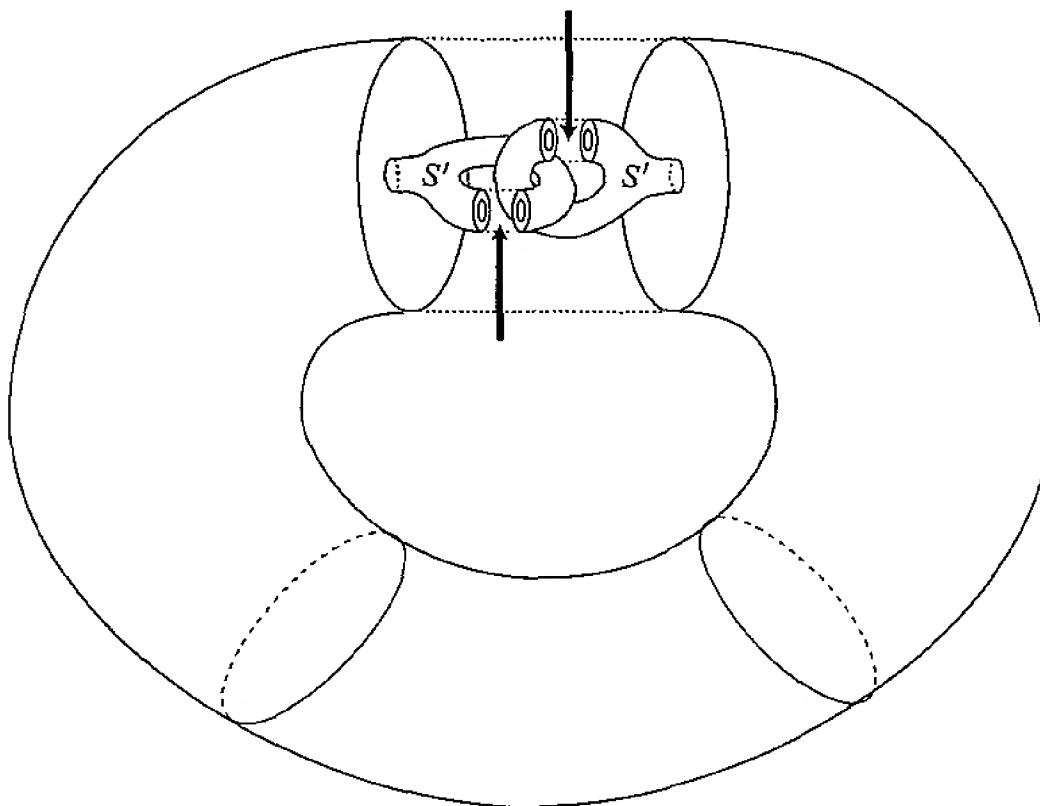
5. (a) If $N \subset M$ is open and \mathcal{A}' consists of all (x, U) in \mathcal{A} with $U \subset N$, show that \mathcal{A}' is maximal for N if \mathcal{A} is maximal for M .
 (b) Show that \mathcal{A}' can also be described as the set of all $(x|V \cap N, V \cap N)$ for (x, V) in \mathcal{A} .
 (c) Show that the inclusion $i: N \rightarrow M$ is C^∞ , and that \mathcal{A}' is the unique atlas with this property.
6. Check that the two projections P_1 and P_2 on S^{n-1} are C^∞ related to the $2n$ homeomorphisms f_i and g_i .
7. (a) If M is a connected C^∞ manifold and $p, q \in M$, then there is a C^∞ curve $c: [0, 1] \rightarrow M$ with $c(0) = p$ and $c(1) = q$.
 (b) It is even possible to choose c to be one-one.
8. (a) Show that $(M_1 \times M_2) \times M_3$ is diffeomorphic to $M_1 \times (M_2 \times M_3)$ and that $M_1 \times M_2$ is diffeomorphic to $M_2 \times M_1$.
 (b) The differentiable structure on $M_1 \times M_2$ makes the “slice” maps

$$\begin{aligned} p_1 &\mapsto (p_1, \bar{p}_2) \\ p_2 &\mapsto (\bar{p}_1, p_2) \end{aligned}$$

of $M_1, M_2 \rightarrow M_1 \times M_2$ differentiable for all $\bar{p}_1 \in M_1, \bar{p}_2 \in M_2$.

- (c) More generally, a map $f: N \rightarrow M_1 \times M_2$ is C^∞ if and only if the compositions $\pi_1 \circ f: N \rightarrow M_1$ and $\pi_2 \circ f: N \rightarrow M_2$ are C^∞ . Moreover, the C^∞ structure we have defined for $M_1 \times M_2$ is the only one with this property.
- (d) If $f_i: N \rightarrow M_i$ are C^∞ ($i = 1, 2$), can one determine the rank of $(f_1, f_2): N \rightarrow M_1 \times M_2$ at p in terms of the ranks of f_i at p ? For $f_i: N_i \rightarrow M_i$, show that $f_1 \times f_2: N_1 \times N_2 \rightarrow M_1 \times M_2$, defined by $f_1 \times f_2(p_1, p_2) = (f_1(p_1), f_2(p_2))$, is C^∞ and determine its rank in terms of the ranks of f_i .
9. Let $g: S^n \rightarrow \mathbb{P}^n$ be the map $p \mapsto [p]$. Show that $f: \mathbb{P}^n \rightarrow M$ is C^∞ if and only if $f \circ g: S^n \rightarrow M$ is C^∞ . Compare the rank of f and the rank of $f \circ g$.
10. (a) If $U \subset \mathbb{R}^n$ is open and $f: U \rightarrow \mathbb{R}$ is locally C^∞ (every point has a neighborhood on which f is C^∞), then f is C^∞ . (Obvious.)
 (b) If $f: \mathbb{H}^n \rightarrow \mathbb{R}$ is locally C^∞ , then f is C^∞ , i.e., f can be extended to a C^∞ function on a neighborhood of \mathbb{H}^n . (Not so obvious.)
11. If $f: \mathbb{H}^n \rightarrow \mathbb{R}$ has two extensions g, h to C^∞ functions in a neighborhood of \mathbb{H}^n , then $D_j g$ and $D_j h$ are the same at points of $\mathbb{R}^{n-1} \times \{0\}$ (so we can speak of $D_j f$ at these points).
12. If M is a C^∞ manifold-with-boundary, then there is a unique C^∞ structure on ∂M such that the inclusion map $i: \partial M \rightarrow M$ is an imbedding.

13. (a) Let $U \subset M^n$ be an open set such that boundary U is an $(n-1)$ -dimensional (differentiable) submanifold. Show that \bar{U} is an n -dimensional manifold-with-boundary. (It is well to bear in mind the following example: if $U = \{x \in \mathbb{R}^n : d(x, 0) < 1 \text{ or } 1 < d(x, 0) < 2\}$, then \bar{U} is a manifold-with-boundary, but $\partial \bar{U} \neq \text{boundary } U$.)
- (b) Consider the figure shown below. This figure may be extended by putting



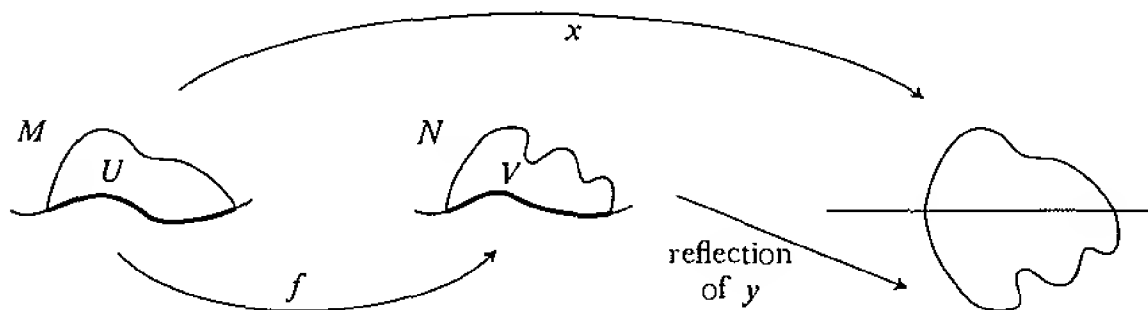
smaller copies of the two parts of S' into the regions indicated by arrows, and then repeating this construction indefinitely. The closure S of the final resulting figure is known as *Alexander's Horned Sphere*. Show that S is homeomorphic to S^2 . (*Hint:* The additional points in the closure are homeomorphic to the Cantor set.) If U is the unbounded component of $\mathbb{R}^3 - S$, then $S = \text{boundary } U$, but \bar{U} is not a 2-dimensional manifold-with-boundary, so part (a) is true only for differentiable submanifolds.

14. (a) There is a map $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that

- (1) $f(x, 0) = (x, 0)$ for all x ,
- (2) $f(x, y) \in \mathbb{H}^2$ for $y \geq 0$,
- (3) $f(x, y) \in \mathbb{R}^2 - \mathbb{H}^2$ for $y < 0$,

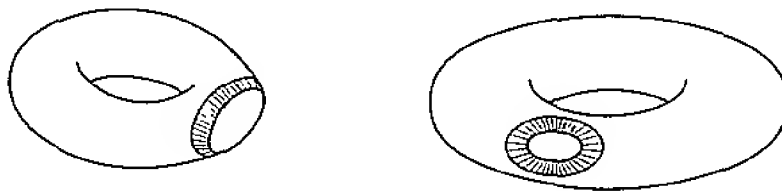
- (4) f restricted to the upper half-plane or the lower half-plane is C^∞ , but f itself is not C^∞ .

(b) Suppose M and N are C^∞ manifolds-with-boundary and $f: \partial M \rightarrow \partial N$ is a diffeomorphism. Let $P = M \cup_f N$ be obtained from the disjoint union of M and N by identifying $x \in \partial M$ with $f(x) \in \partial N$. If (x, U) is a coordinate system around $p \in \partial M$ and (y, V) a coordinate system around $f(p)$, with $f(U \cap \partial M) = V \cap \partial N$, and $(y \circ f)|_{U \cap \partial M} = x|_{U \cap \partial M}$, we can define a homeomorphism from $U \cup V \subset P$ to \mathbb{R}^n by sending U to \mathbb{H}^n by x and V to the lower half-plane by the reflection of y . Show that this procedure does *not*



define a C^∞ structure on P .

(c) Now suppose that there is a neighborhood U of ∂M in M and a diffeomorphism $\alpha: U \rightarrow \partial M \times [0, 1)$, such that $\alpha(p) = (p, 0)$ for all $p \in \partial M$, and a similar diffeomorphism $\beta: V \rightarrow \partial N \times [0, 1)$. (We will be able to prove later that such diffeomorphisms always exist). Show that there is a unique C^∞ structure



on P such that the inclusions of M and N are C^∞ and such that the map from $U \cup V$ to $\partial M \times (-1, 1)$ induced by α and β is a diffeomorphism.

(d) By using two different pairs (α, β) , define two different C^∞ structures on \mathbb{R}^2 , considered as the union of two copies of \mathbb{H}^2 with corresponding points on $\partial \mathbb{H}^2$ identified. Show that the resulting C^∞ manifolds are diffeomorphic, but that the diffeomorphism *cannot* be chosen arbitrarily close to the identity map.

15. (a) Find a C^∞ structure on $\mathbb{H}^1 \times \mathbb{H}^1$ which makes the inclusion into \mathbb{R}^2 a C^∞ map. Can the inclusion be an imbedding? Are the projections on each factor C^∞ maps?

(b) If M and N are manifolds-with-boundary, construct a C^∞ structure on $M \times N$ such that all the "slice maps" (defined in Problem 8) are C^∞ .

16. Show that the function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

is C^∞ (the formula e^{-1/x^2} is used just to get a function which is > 0 for $x < 0$, and $e^{-1/|x|}$ could be used just as well).

17. Lemma 2 (as addended by the proof of Theorem 15) shows that if C_1 and C_2 are disjoint closed subsets of M , then there is a C^∞ function $f: M \rightarrow [0, 1]$ such that $C_1 \subset f^{-1}(0)$ and $C_2 \subset f^{-1}(1)$. Actually, we can even find f with $C_1 = f^{-1}(0)$ and $C_2 = f^{-1}(1)$. The proof turns out to be quite easy, once you know the trick.

(a) It suffices to find, for any closed $C \subset M$, a C^∞ function f with $C = f^{-1}(0)$.

(b) Let $\{U_i\}$ be a countable cover of $M - C$, where each U_i is of the form

$$U_i = x^{-1}(\{a \in \mathbb{R}^n : |a| < 1\})$$

for some coordinate system x taking an open subset of $M - C$ onto \mathbb{R}^n . Let $f_i: M \rightarrow [0, 1]$ be a C^∞ function with $f_i > 0$ on U_i and $f_i = 0$ on $M - U_i$. Functions like

$$\frac{\partial f_i}{\partial x^j}, \frac{\partial^2 f_i}{\partial x^j \partial x^k}, \dots$$

will be called mixed partials of f_i , of order $1, 2, \dots$. Let

$$\alpha_i = \sup \text{ of all mixed partials of } f_i, \dots, f_i \text{ of all orders } \leq i.$$

Show that

$$f = \sum_{i=1}^{\infty} \frac{f_i}{\alpha_i 2^i}$$

is C^∞ , and $C = f^{-1}(0)$.

18. Consider the coordinate system (y^1, y^2) for \mathbb{R}^2 defined by

$$\begin{aligned} y^1(a, b) &= a \\ y^2(a, b) &= a + b. \end{aligned}$$

- (a) Compute $\partial f / \partial y^1(a, b)$ from the definition.
 (b) Also compute it from Proposition 3 (to find $\partial I^i / \partial y^j$, write each I^i in terms of y^1 and y^2).

Notice that $\partial f / \partial y^1 \neq \partial f / \partial I^1$ even though $y^1 = I^1$; the operator $\partial / \partial y^i$ depends on y and i , not just on y^i .

19. Compute the “Laplacian”

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

in terms of polar coordinates. (First compute $\partial / \partial x$ in terms of $\partial / \partial r$ and $\partial / \partial \theta$; then compute $\partial^2 / \partial x^2$ from this). *Answer:* $\frac{1}{r} \left[\frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \right) + \frac{\partial}{\partial \theta} \left(\frac{1}{r} \frac{\partial}{\partial \theta} \right) \right]$.

20. If $f: M^n \rightarrow N^m$ is C^1 and $m > n$, then $f(M)$ has measure 0 (provided that M has only countably many components).

21. The following pictures show, for $n = 1, 2$, and 3, a subdivision of $[0, 1] \times [0, 1]$ into 2^{2n} squares, $A_{n,1}, \dots, A_{n,2^{2n}}$; square $A_{n,k}$ is labeled simply k . The numbering is determined by the following conditions:

- (a) The lower left square is $A_{n,1}$.
- (b) The upper left square is $A_{n,2^{2n}}$.
- (c) Squares $A_{n,k}$ and $A_{n,k+1}$ have a common side.
- (d) Squares $A_{n,4l+1}, A_{n,4l+2}, A_{n,4l+3}, A_{n,4l+4}$ are contained in $A_{n-1,l+1}$.

4	3
1	2

16	13	12	11
15	14	9	10
2	3	8	7
1	4	5	6

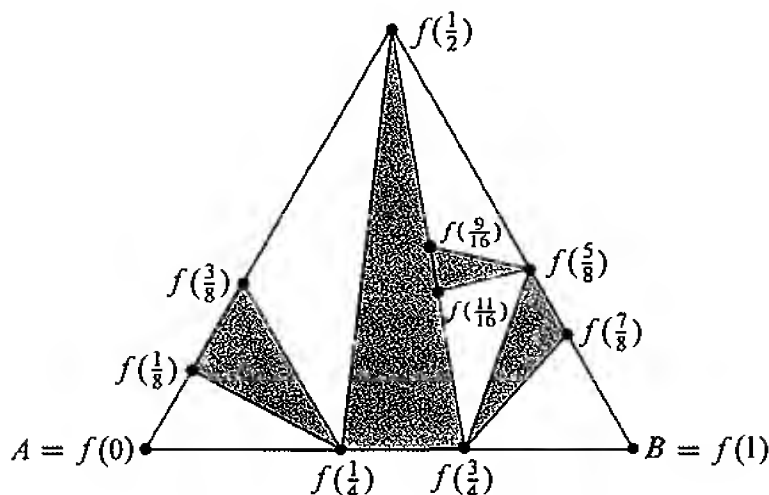
6	7	10	11				
5	8	9	12				
4	3		13				
1	2						

Define $f: [0, 1] \rightarrow [0, 1] \times [0, 1]$ by the condition

$$f(t) \in A_{n,k} \quad \text{for all} \quad \frac{k-1}{2^{2n}} \leq t \leq \frac{k}{2^{2n}}.$$

Show that f is continuous, onto $[0, 1] \times [0, 1]$, and not one-one.

22. For $p/2^n \in [0, 1]$, define $f(p/2^n) \in \mathbb{R}^2$ as shown below.



(a) Show that f is uniformly continuous, so that it has a continuous extension $g: [0, 1] \rightarrow \mathbb{R}^2$. Show that g is one-one, and that its image will not have measure 0 if the shaded triangles are chosen correctly.

(b) Consider the homeomorphic image of S^1 obtained by adding, below the image of g , a semi-circle with diameter the line segment AB . What does the inside of this curve look like?

23. Let $c: [0, 1] \rightarrow \mathbb{R}^n$ be continuous. For each partition $P = \{t_0, \dots, t_k\}$ of $[0, 1]$, define

$$\ell(c, P) = \sum_{i=1}^k d(c(t_i), c(t_{i-1})).$$

The curve c is rectifiable if $\{\ell(c, P)\}$ is bounded above (with length equal to $\sup\{\ell(c, P)\}$). Show that the image of a rectifiable curve has measure 0.

24. (a) If M is a C^∞ manifold, a set $M_1 \subset M$ can be made into a k -dimensional submanifold of M if and only if around each point in M_1 there is a coordinate system (x, U) on M such that $M_1 \cap U = \{p: x^{k+1}(p) = \dots = x^n(p) = 0\}$.

(b) The subset M_1 can be made into a closed submanifold if and only if such coordinate systems exist around every point of M .

25. The set $\{(x, |x|): x \in \mathbb{R}\}$ is not the image of any immersion of \mathbb{R} into \mathbb{R}^2 .

26. (a) If $U \subset \mathbb{R}^k$ is open and $f: U \rightarrow \mathbb{R}^{n-k}$ is C^∞ , then the graph of $f = \{(p, f(p)) \in \mathbb{R}^n: p \in U\}$ is a submanifold of \mathbb{R}^n .

(b) Every submanifold of \mathbb{R}^n is locally of this form, after renumbering coordinates. (Neither Theorem 9 nor 10 is quite strong enough. You will need

the implicit function theorem (*Calculus on Manifolds*, pg. 41). Theorem 10 is essentially Theorem 2-13 of *Calculus on Manifolds*; comparison with the implicit function theorem will show how some information has been allowed to escape.)

27. (a) An immersion from one n -manifold to another is an open map (the image of an open set is open).

(b) If M and N are n -manifolds with M compact and N connected, and $f: M \rightarrow N$ is an immersion, then f is onto.

28. Prove Proposition 12: If $f: M^n \rightarrow N$ has constant rank k on a neighborhood of $f^{-1}(y)$, then $f^{-1}(y)$ is a (closed) submanifold of M of dimension $n - k$ (or is empty).

29. Let $f: \mathbb{P}^2 \rightarrow \mathbb{R}^3$ be the map

$$g([x, y, z]) = (yz, xz, xy)$$

defined in Chapter 1, whose image is the Steiner surface. Show that g fails to be an immersion at 6 points (the image points are the points at distance $\pm 1/2$ on each axis). There is a way of immersing \mathbb{P}^2 in \mathbb{R}^3 , known as Boy's Surface. See Hilbert and Cohn-Vossen, *Geometry and the Imagination*, pp. 317–321.

30. A continuous function $f: X \rightarrow Y$ is proper if $f^{-1}(C)$ is compact for every compact $C \subset Y$. The limit set $L(f)$ of f is the set of all $y \in Y$ such that $y = \lim f(x_n)$ for some sequence $x_1, x_2, x_3, \dots \in X$ with no convergent subsequence.

(a) $L(f) = \emptyset$ if and only if f is proper.

(b) $f(X) \subset Y$ is closed if and only if $L(f) \subset f(X)$.

(c) There is a continuous $f: \mathbb{R} \rightarrow \mathbb{R}^2$ with $f(\mathbb{R})$ closed, but $L(f) \neq \emptyset$.

(d) A one-one continuous function $f: X \rightarrow Y$ is a homeomorphism (onto its image) if and only if $L(f) \cap f(Y) = \emptyset$.

(e) A submanifold $M_1 \subset M$ is a closed submanifold if and only if the inclusion map $i: M_1 \rightarrow M$ is proper.

(f) If M is a manifold, there is a proper map $f: M \rightarrow \mathbb{R}$; the function f can be made C^∞ if M is a C^∞ manifold.

31. (a) Find a cover of $[0, 1]$ which is not locally finite but which is “point-finite”: every point of $[0, 1]$ is in only finitely many members of the cover.

(b) Prove the Shrinking Lemma when the cover \mathcal{O} is point-finite and countable (notice that local-finiteness is not really used).

(c) Prove the Shrinking Lemma when \mathcal{O} is a (not necessarily countable) point-finite cover of any space. (You will need Zorn's Lemma; consider collections \mathcal{C}

of pairs (U, U') where $U \in \mathcal{O}$, $U' \subset U$, and the union of all U' for $(U, U') \in \mathcal{C}$, together with all other $U \in \mathcal{O}$ covers the space.)

32. (a) If $M_1 \subset M$ is a closed submanifold, $U \supset M_1$ is any neighborhood, and $f: M_1 \rightarrow \mathbb{R}$ is C^∞ , then there is a C^∞ function $\tilde{f}: M \rightarrow \mathbb{R}$ with $\tilde{f} = f$ on M_1 , and with support $\tilde{f} \subset U$.
 (b) This is false if $M = \mathbb{R}$ and $M_1 = (0, 1)$.
 (c) This is false if \mathbb{R} is replaced by a disconnected manifold N .

Remark: It is also false if $M = \mathbb{R}^2$, $M_1 = N = S^1$, and $f = \text{identity}$; in fact, in this case, f has no continuous extension to a map from \mathbb{R}^2 to S^1 , but the proof requires some topology. However, f can always be extended to a C^∞ function in a neighborhood of M_1 (extend locally, and use partitions of unity).

33. (a) The set of all non-singular $n \times n$ matrices with real entries is called $GL(n, \mathbb{R})$, the **general linear group**. It is a C^∞ manifold, since it is an open subset of \mathbb{R}^{n^2} . The **special linear group** $SL(n, \mathbb{R})$, or **unimodular group**, is the subgroup of all matrices with $\det = 1$. Using the formula for $D(\det)$ in *Calculus on Manifolds*, pg. 24, show that $SL(n, \mathbb{R})$ is a closed submanifold of $GL(n, \mathbb{R})$ of dimension $n^2 - 1$.
 (b) The symmetric $n \times n$ matrices may be thought of as $\mathbb{R}^{n(n+1)/2}$. Define $\psi: GL(n, \mathbb{R}) \rightarrow (\text{symmetric matrices})$ by $\psi(A) = A \cdot A^t$, where A^t is the transpose of A . The subgroup $\psi^{-1}(I)$ of $GL(n, \mathbb{R})$ is called the **orthogonal group** $O(n)$. Show that $A \in O(n)$ if and only if the rows [or columns] of A are orthonormal.
 (c) Show that $O(n)$ is compact.
 (d) For any $A \in GL(n, \mathbb{R})$, define $R_A: GL(n, \mathbb{R}) \rightarrow GL(n, \mathbb{R})$ by $R_A(B) = BA$. Show that R_A is a diffeomorphism, and that $\psi \circ R_A = \psi$ for all $A \in O(n)$. By applying the chain rule, show that for $A \in O(n)$ the matrix

$$\left(\frac{\partial \psi^{ij}}{\partial x^{kl}}(A) \right) \text{ has the same rank as } \left(\frac{\partial \psi^{ij}}{\partial x^{kl}}(I) \right).$$

(Here x^{kl} are the coordinate functions in \mathbb{R}^{n^2} , and ψ^{ij} the $n(n+1)/2$ component functions of ψ .) Conclude from Proposition 12 that $O(n)$ is a submanifold of $GL(n, \mathbb{R})$.

- (e) Using the formula

$$\psi^{ij}(A) = \sum_k a_{ik} a_{jk} \quad (A = (a_{ij})),$$

show that

$$\frac{\partial \psi^{ij}}{\partial x^{kl}}(A) = \begin{cases} a_{jl} & k = i \neq j \\ a_{il} & k = j \neq i \\ 2a_{il} & k = i = j \\ 0 & \text{otherwise.} \end{cases}$$

Show that the rank of this matrix is $n(n+1)/2$ at I (and hence at A for all $A \in O(n)$.) Conclude that $O(n)$ has dimension $n(n-1)/2$.

(f) Show that $\det A = \pm 1$ for all $A \in O(n)$. The group $O(n) \cap SL(n, \mathbb{R})$ is called the **special orthogonal group** $SO(n)$, or the **rotation group** $R(n)$.

34. Let $M(m, n)$ denote the set of all $m \times n$ matrices, and $M(m, n; k)$ the set of all $m \times n$ matrices of rank k .

(a) For every $X_0 \in M(m, n; k)$ there are permutation matrices P and Q such that

$$PX_0Q = \begin{pmatrix} A_0 & B_0 \\ C_0 & D_0 \end{pmatrix}, \quad \text{where } A_0 \text{ is } k \times k \text{ and non-singular.}$$

(b) There is some $\varepsilon > 0$ such that A is non-singular whenever all entries of $A - A_0$ are $< \varepsilon$.

(c) If

$$PXQ = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where the entries of $A - A_0$ are $< \varepsilon$, then X has rank k if and only if $D = CA^{-1}B$. *Hint:* If I_k denotes the $k \times k$ identity matrix, then

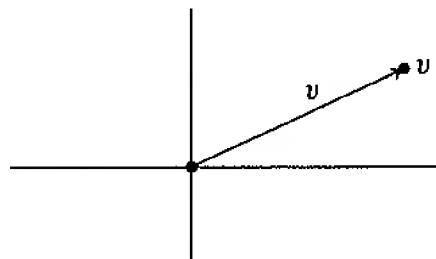
$$\begin{pmatrix} I_k & 0 \\ X & I_{p-k} \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & B \\ XA + C & XB + D \end{pmatrix}.$$

(d) $M(m, n; k) \subset M(m, n)$ is a submanifold of dimension $k(m+n-k)$ for all $k \leq m, n$.

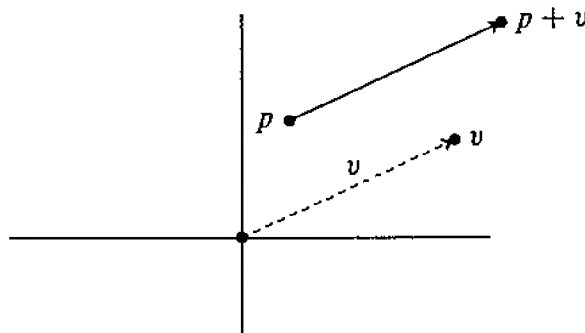
CHAPTER 3

THE TANGENT BUNDLE

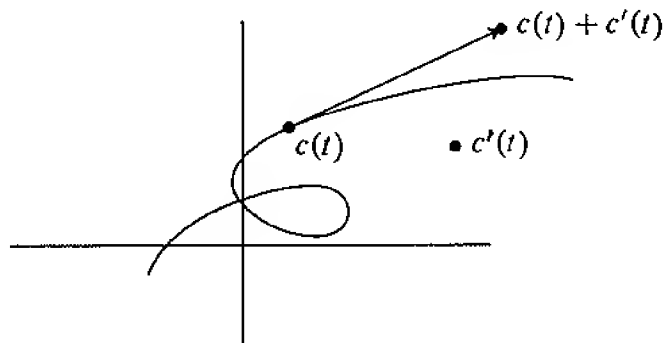
A point $v \in \mathbb{R}^n$ is frequently pictured as an arrow from 0 to v . But there are many situations where we would like to picture this same arrow as starting



at a different point $p \in \mathbb{R}^n$:

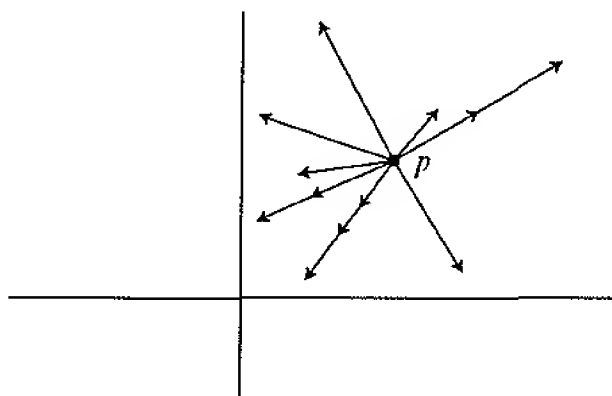


For example, suppose $c: \mathbb{R} \rightarrow \mathbb{R}^n$ is a differentiable curve. Then $c'(t) = (c^1(t), \dots, c^n(t))$ is just a point of \mathbb{R}^n , but the line between $c(t)$ and $c(t) + c'(t)$ is tangent to the curve, and the “velocity vector” or “tangent vector” $c'(t)$ of the curve c is customarily pictured as the arrow from $c(t)$ to $c(t) + c'(t)$.

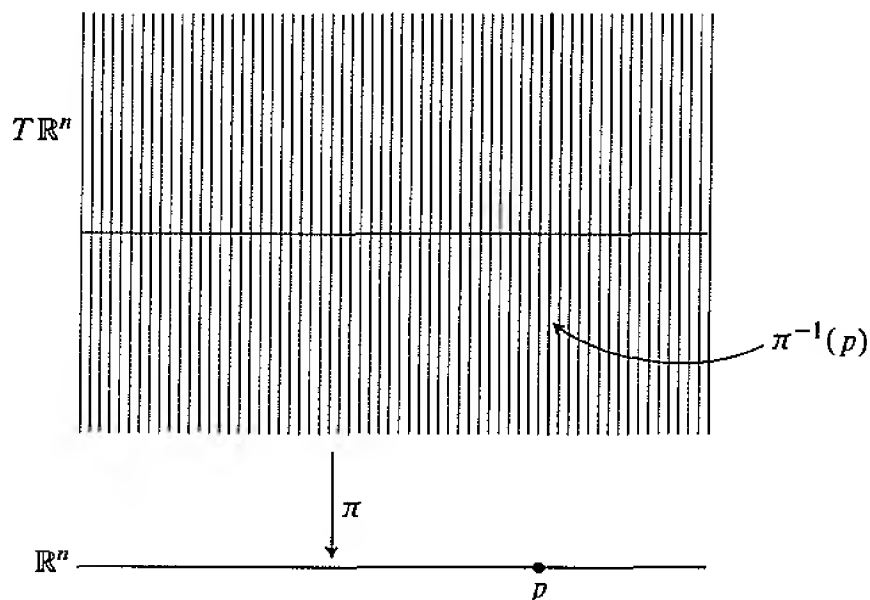


To give this picture mathematical substance, we simply describe the “arrow” from p to $p + v$ by the pair (p, v) . The set of all such pairs is just $\mathbb{R}^n \times \mathbb{R}^n$, which we will also denote by $T\mathbb{R}^n$, the “tangent space of \mathbb{R}^n ”; elements of $T\mathbb{R}^n$ are called “tangent vectors” of \mathbb{R}^n . We will often denote $(p, v) \in T\mathbb{R}^n$ by v_p (“the vector v at p ”); in conformity with this notation, we will denote the set of all (p, v) for $v \in \mathbb{R}^n$ by \mathbb{R}^n_p . At times, it is more convenient to denote a member of $T\mathbb{R}^n$ by a single letter, like v . To recover the first member of a pair $v \in T\mathbb{R}^n$, we define the “projection” map $\pi: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $\pi(a, b) = a$. For any tangent vector v , the point $\pi(v)$ is “where it’s at”.

The set $\pi^{-1}(p)$ may be pictured as all arrows starting at p . Alternately,



it can be pictured more geometrically as a particular subset of $\mathbb{R}^n \times \mathbb{R}^n$, the one visualizable case occurring when $n = 1$. This picture gives rise to some



terminology—we call $\pi^{-1}(p)$ the fibre over p . This fibre can be made into a

vector space in an obvious way: we define

$$(p, v) \oplus (p, w) = (p, v + w) \\ a \bullet (p, v) = (p, a \cdot v).$$

(The operations \oplus and \bullet should really be thought of as defined on

$$\bigcup_{p \in \mathbb{R}^n} \pi^{-1}(p) \times \pi^{-1}(p), \quad \text{and} \quad \mathbb{R} \times T\mathbb{R}^n, \quad \text{respectively.}$$

Usually we will just use ordinary $+$ and \cdot instead of \oplus and \bullet .)

If $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a differentiable map, and $p \in \mathbb{R}^n$, then the linear transformation $Df(p): \mathbb{R}^n \rightarrow \mathbb{R}^m$ may be used to produce a linear map from $\mathbb{R}^n_p \rightarrow \mathbb{R}^m_{f(p)}$ defined by

$$v_p \mapsto [Df(p)(v)]_{f(p)}.$$

This map, whose apparently anomalous features will soon be justified, is denoted by f_{*p} ; the symbol f_* denotes the map $f_*: T\mathbb{R}^n \rightarrow T\mathbb{R}^m$ which is the union of all f_{*p} . Since $f_{*p}(v)$ is defined to be a vector $\in \mathbb{R}^m_{f(p)}$, the following diagram “commutes” (the two possible compositions from $T\mathbb{R}^n$ to \mathbb{R}^m are equal),

$$\begin{array}{ccc} T\mathbb{R}^n & \xrightarrow{f_*} & T\mathbb{R}^m \\ \pi \downarrow & & \downarrow \pi \\ \mathbb{R}^n & \xrightarrow{f} & \mathbb{R}^m \end{array} \quad \pi \circ f_* = f \circ \pi.$$

Thus, f_* has the map f , as well as all maps $Df(p)$, built into it.

This is not the only reason for defining f_* in this particular way, however. Suppose that $g: \mathbb{R}^m \rightarrow \mathbb{R}^k$ is another differentiable function, so that, by the chain rule,

$$(1) \quad D(g \circ f)(p) = Dg(f(p)) \circ Df(p).$$

By our definition,

$$g_*([Df(p)(v)]_{f(p)}) = (Dg(f(p))(Df(p)(v)))_{g(f(p))}.$$

This looks horribly complicated, but, using (1), it can be written

$$g_*(f_*(v_p)) = (g \circ f)_*(v_p);$$

thus we have

$$g_* \circ f_* = (g \circ f)_*.$$

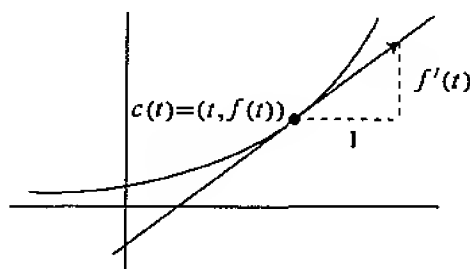
This relation would clearly fall apart completely if $f_*(v_p)$ were not in $\mathbb{R}^m_{f(p)}$; with our present definition of f_* , it is merely an elegant restatement of the chain rule.

Henceforth, we will state almost all concepts about Jacobian matrices, like rank or singularity, in terms of f_* , rather than Df . The “tangent vector” of a curve $c: \mathbb{R} \rightarrow \mathbb{R}^n$ can be defined in terms of this concept, also. The tangent vector of c at t may be defined as

$$c'(t)_{c(t)} \in \mathbb{R}^n_{c(t)}.$$

[If c happens to be of the form

$$c(t) = (t, f(t)) \text{ for } f: \mathbb{R} \rightarrow \mathbb{R}$$



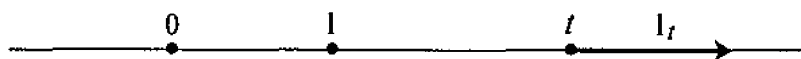
then

$$c'(t)_{c(t)} = (1, f'(t))_{c(t)};$$

this vector lies along the tangent line to the graph of f at $(t, f(t))$.] Notice that the tangent vector of c at t is the same as

$$c_*(1_t) = [Dc(t)(1)]_{c(t)} = (c^{1'}(t), \dots, c^{n'}(t))_{c(t)},$$

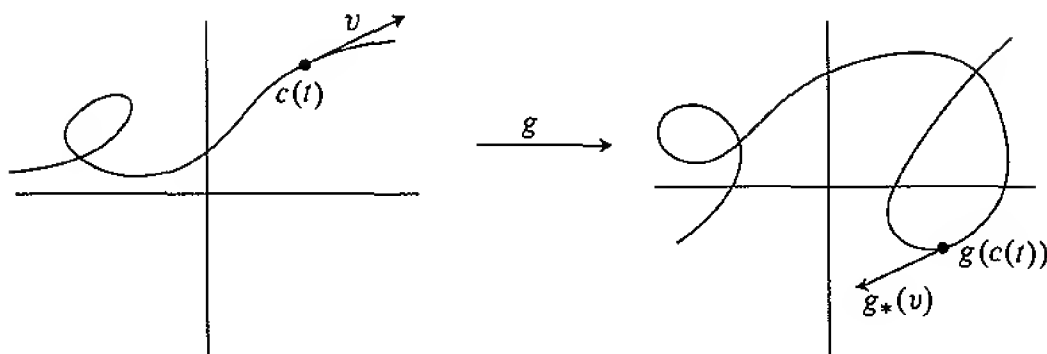
where $1_t = (t, 1)$ is the “unit” tangent vector of \mathbb{R} at t .



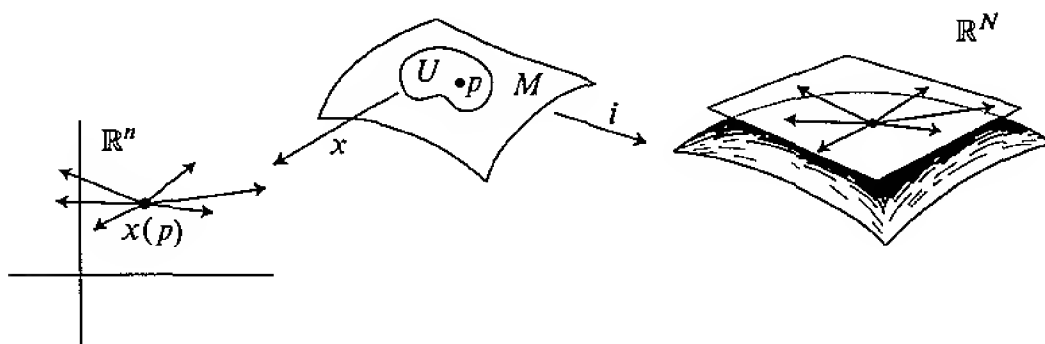
If $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable, then $g \circ c$ is a curve in \mathbb{R}^m . The tangent

vector of $g \circ c$ at t is

$$\begin{aligned} (g \circ c)_*(l_t) &= g_*(c_*(l_t)) \\ &= g_*(\text{tangent vector of } c \text{ at } t). \end{aligned}$$



Consider now an n -dimensional manifold M and an imbedding $i: M \rightarrow \mathbb{R}^N$. Suppose we take a coordinate system (x, U) around p . Then $i \circ x^{-1}$ is a map from \mathbb{R}^n to \mathbb{R}^N with rank n . Consequently, $(i \circ x^{-1})_*(\mathbb{R}^n_{x(p)})$ is an n -dimensional subspace of $\mathbb{R}^N_{i(p)}$. This subspace doesn't depend on the coordinate



system x , for if y is another coordinate system, then

$$\begin{aligned} (i \circ y^{-1})_* &= (i \circ x^{-1} \circ x \circ y^{-1})_* \\ &= (i \circ x^{-1})_* \circ (x \circ y^{-1})_* \end{aligned}$$

and

$$(x \circ y^{-1})_{*y(p)}: \mathbb{R}^n_{y(p)} \rightarrow \mathbb{R}^n_{x(p)}$$

is an isomorphism (with inverse $(y \circ x^{-1})_{*x(p)}$).

There is another way to see this, which justifies the picture we have drawn. If $c: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$ is a curve with $c(0) = x(p)$, then $\alpha = i \circ x^{-1} \circ c$ is a curve in \mathbb{R}^N which lies in $i(M)$, and every differentiable curve in $i(M)$ is of this



form (Proof?). Now

$$\alpha_*(1_0) = (i \circ x^{-1})_* \circ c_*(1_0),$$

so the tangent vector of every α is in $(i \circ x^{-1})_*(\mathbb{R}^n_{x(p)})$. Moreover, every vector in this subspace is the tangent vector of some α , since every vector in $\mathbb{R}^n_{x(p)}$ is the tangent vector of some curve c . Thus, our n -dimensional subspace is just the set of all tangent vectors at $i(p)$ to differentiable curves in $i(M)$. We will denote this n -dimensional subspace by $(M, i)_p$.

We now want to look at the (disjoint) union

$$T(M, i) = \bigcup_{p \in M} (M, i)_p \subset i(M) \times \mathbb{R}^N \subset T\mathbb{R}^N.$$

We can define a “projection” map

$$\pi: T(M, i) \rightarrow M$$

by

$$\pi(v) = p \quad \text{if} \quad v \in (M, i)_p.$$

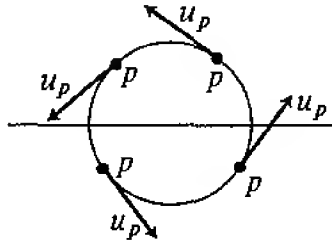
As in the case of $T\mathbb{R}^n$, each “fibre” $\pi^{-1}(p)$ has a vector space structure also. Beyond this we have to look a little more carefully at some specific examples.

Consider first the manifold $M = S^1$ and the inclusion $i: S^1 \rightarrow \mathbb{R}^2$. The curve $c(\theta) = (\cos \theta, \sin \theta)$ passes through every point of S^1 , and

$$c'(\theta) = (-\sin \theta, \cos \theta) \neq 0.$$

For each $p = (\cos \theta, \sin \theta) \in S^1$, let $u_p = (-\sin \theta, \cos \theta)_p$ (it clearly doesn't matter which of the infinitely many possible θ 's we choose). Then $(S^1, i)_p$ con-

sists of all multiples of the vector u_p . We can therefore define a homeomorphism



$f_1: T(S^1, i) \rightarrow S^1 \times \mathbb{R}^1$ by $f_1(\lambda u_p) = (p, \lambda)$, which makes the following diagram commute.

$$\begin{array}{ccc}
 T(S^1, i) & \xrightarrow{f_1} & S^1 \times \mathbb{R}^1 \\
 \searrow \pi & & \swarrow \pi' \\
 & S^1 &
 \end{array}
 \quad [\pi'(a, b) = a]$$

If we define the “fibres” of π' to be the sets $\pi'^{-1}(p)$, then each fibre has a vector space structure in a natural way. Commutativity of the diagram means that f_1 takes fibres into fibres; clearly f_1 restricted to a fibre is a linear isomorphism onto the image.

Now consider the manifold $M = S^2$ and the inclusion $i: S^2 \subset \mathbb{R}^3$. In this case there is no map $f_2: T(S^2, i) \rightarrow S^2 \times \mathbb{R}^2$ with the properties of the map f_1 . If there were, then, for a fixed vector $v \neq 0$ in \mathbb{R}^2 , the set of vectors

$$\{f_2^{-1}(v_p) : p \in S^2\}$$

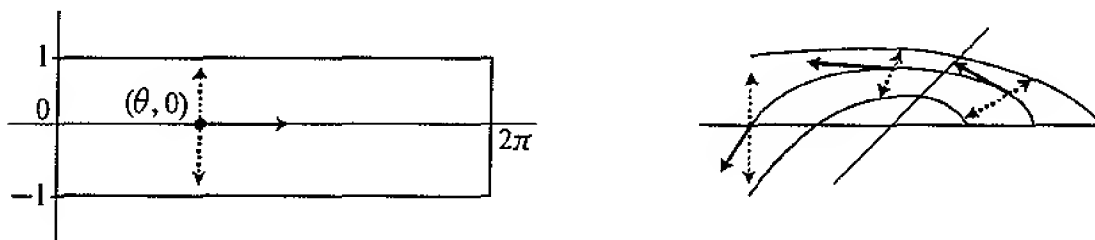
would be a collection of non-zero tangent vectors, one at each point of S^2 , which varied continuously. It is a well-known (hard) theorem of topology that this is impossible (you can't comb the hair on a sphere).



There is another example where we can prove that no appropriate homeomorphism $T(M, i) \rightarrow M \times \mathbb{R}^2$ exists, without appealing to a hard theorem of topology. The map i will just be the inclusion $M \rightarrow \mathbb{R}^3$ where M is a Möbius strip, to be precise, the particular subset of \mathbb{R}^3 defined in Chapter 1— M is the image of the map $f: [0, 2\pi] \times (-1, 1) \rightarrow \mathbb{R}^3$ defined by

$$f(\theta, t) = \left(2 \cos \theta + t \cos \frac{\theta}{2} \cos \theta, 2 \sin \theta + t \cos \frac{\theta}{2} \sin \theta, t \sin \frac{\theta}{2} \right).$$

At each point $p = (2 \cos \theta, 2 \sin \theta, 0)$ of M , the vector



$$v_p = (-2 \sin \theta, 2 \cos \theta, 0)_p = f_*((1, 0)_{(\theta, 0)})$$

is a tangent vector. The same is true for all multiples of $f_*((0, 1)_{(\theta, 0)})$, shown as dashed arrows in the picture. Notice that

$$\begin{aligned} f_*((0, 1)_{(0, 0)}) &= [Df(0, 0)(0, 1)]_{(2, 0, 0)} \\ &= \left[\frac{\partial f}{\partial t}(0, 0) \right]_{(2, 0, 0)} = (1, 0, 0)_{(2, 0, 0)}, \end{aligned}$$

while

$$f_*((0, 1)_{(2\pi, 0)}) = \left[\frac{\partial f}{\partial t}(2\pi, 0) \right]_{(2, 0, 0)} = (-1, 0, 0)_{(2, 0, 0)}.$$

This means that we can never pick non-zero dashed vectors *continuously* on the set of all points $(2 \cos \theta, 2 \sin \theta, 0)$: If we could, then each vector would be

$$f_*((0, \lambda(\theta))_{(\theta, 0)})$$

for some continuous function $\lambda: [0, 2\pi] \rightarrow \mathbb{R}$. This function would have to be non-zero everywhere and also satisfy $\lambda(2\pi) = -\lambda(0)$, which it can't (by an easy theorem of topology). The impossibility of choosing non-zero dashed vectors continuously clearly shows that there is no way to map $T(M, i)$, fibre by fibre,

homeomorphically onto $M \times \mathbb{R}^2$. We thus have another case where $T(M, i)$ does not “look like” a product $M \times \mathbb{R}^n$.

For any imbedding $i: M \rightarrow \mathbb{R}^N$, however, the structure of $T(M, i)$ is always simple *locally*: if (x, U) is a coordinate system on M , then $\pi^{-1}(U)$, the part of $T(M, i)$ over U , can always be mapped, fibre by fibre, homeomorphically onto $U \times \mathbb{R}^n$. In fact, for each $p \in U$, the fibre

$$(M, i)_p \text{ equals } (i \circ x^{-1})_{*x(p)} (\mathbb{R}^n_{x(p)}) = m_p (\mathbb{R}^n_{x(p)}),$$

where the abbreviation m_p has been introduced temporarily; we can therefore define

$$f: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$$

by

$$f(m_p(v_{x(p)})) = (p, v).$$

In standard jargon, $T(M, i)$ is “locally trivial”. This additional feature qualifies $T(M, i)$ to be included among an extremely important class of structures:

An n -dimensional vector bundle (or n -plane bundle) is a five-tuple

$$\xi = (E, \pi, B, \oplus, \odot),$$

where

- (1) E and B are spaces (the “total space” and “base space” of ξ , respectively),
- (2) $\pi: E \rightarrow B$ is a continuous map *onto* B ,
- (3) \oplus and \odot are maps

$$\oplus: \bigcup_{p \in B} \pi^{-1}(p) \times \pi^{-1}(p) \rightarrow E, \quad \odot: \mathbb{R} \times E \rightarrow E,$$

with $\oplus(\pi^{-1}(p) \times \pi^{-1}(p)) \subset \pi^{-1}(p)$ and $\odot(\mathbb{R} \times \pi^{-1}(p)) \subset \pi^{-1}(p)$, which make each fibre $\pi^{-1}(p)$ into an n -dimensional vector space over \mathbb{R} ,

such that the following “local triviality” condition is satisfied:

For each $p \in B$, there is a neighborhood U of p and a homeomorphism $t: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$ which is a vector space isomorphism from each $\pi^{-1}(q)$ onto $q \times \mathbb{R}^n$, for all $q \in U$.

Because this local triviality condition really is a local condition, each bundle $\xi = (E, \pi, B, \oplus, \odot)$ automatically gives rise to a bundle $\xi|A$ over any subset $A \subset B$; to be precise,

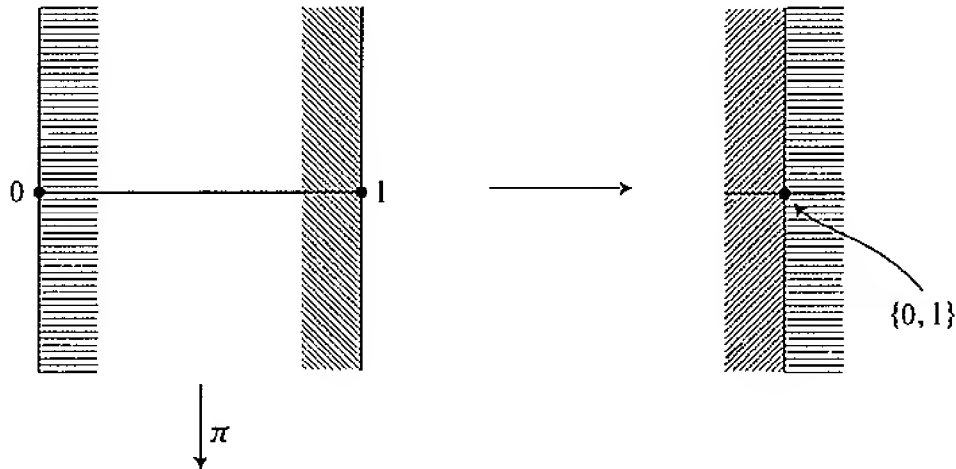
$$\xi|A = \left(\pi^{-1}(A), \pi|_{\pi^{-1}(A)}, A, \oplus|_{\bigcup_{p \in A} \pi^{-1}(p) \times \pi^{-1}(p)}, \odot|_{\mathbb{R} \times \pi^{-1}(A)} \right).$$

Notation as cumbersome as all this invites abuse, and we shall usually refer simply to a bundle $\pi: E \rightarrow B$, or even denote the bundle by E alone. For vectors $v, w \in \pi^{-1}(p)$ and $a \in \mathbb{R}$, we will denote $\oplus(v, w)$ and $\odot(a, v)$ by $v + w$, and $a \cdot v$ or av , respectively.

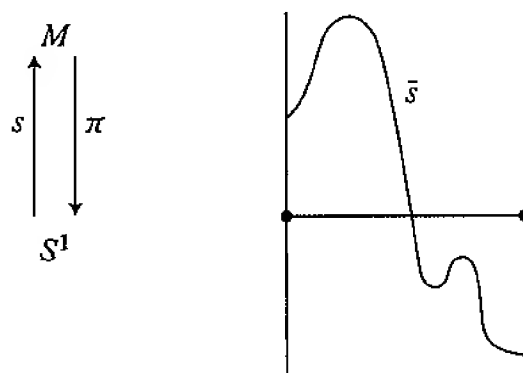
The simplest example of an n -plane bundle is just $X \times \mathbb{R}^n$ with $\pi: X \times \mathbb{R}^n \rightarrow X$ the projection on the first factor, and the obvious vector space structure on each fibre. This is called the trivial n -plane bundle over X and will be denoted by $\varepsilon^n(X)$. The “tangent bundle” $T\mathbb{R}^n$ is just $\varepsilon^n(\mathbb{R}^n)$.

The bundle $T(S^1, i)$ considered before is equivalent to $\varepsilon^1(S^1)$. Equivalence is here a technical term: Two vector bundles $\xi_1 = \pi_1: E_1 \rightarrow B$ and $\xi_2 = \pi_2: E_2 \rightarrow B$ are equivalent ($\xi_1 \simeq \xi_2$) if there is a homeomorphism $h: E_1 \rightarrow E_2$ which takes each fibre $\pi_1^{-1}(p)$ isomorphically onto $\pi_2^{-1}(p)$. The map h is called an equivalence. A bundle equivalent to $\varepsilon^n(B)$ is called **trivial**. (The local triviality condition for a bundle ξ just says that $\xi|U$ is trivial for some neighborhood U of p .)

The bundles $T(S^2, i)$ and $T(M, i)$ are not trivial, but there is an even simpler example of a non-trivial bundle. The Möbius strip *itself* (not $T(M, i)$) can be considered as a 1-dimensional vector bundle over S^1 , for M can be obtained from $[0, 1] \times \mathbb{R}$ by identifying $(0, a)$ with $(1, -a)$, while S^1 can be obtained from



$[0, 1]$ by identifying 0 with 1; the map π is defined by $\pi(t, a) = t$ for $0 < t < 1$ and $\pi(\{(0, a), (1, -a)\}) = \{0, 1\}$. The diagram above illustrates local triviality near the point $\{0, 1\}$ of S^1 . Suppose that $s: S^1 \rightarrow M$ is a continuous function with $\pi \circ s = \text{identity of } M$ (such a function is called a section). Such a map



corresponds to a continuous function $\bar{s}: [0, 1] \rightarrow \mathbb{R}$ with $\bar{s}(0) = -\bar{s}(1)$. Since \bar{s} must be 0 somewhere, the section s must be 0 somewhere (that is, $s(\theta) \in \pi^{-1}(\theta)$ must be the 0 vector for some $\theta \in S^1$). This surely shows that M is not a trivial bundle.

An equivalence is obviously the analogue of an isomorphism. The analogue of a homomorphism is the following.* A bundle map from ξ_1 to ξ_2 is a pair of continuous maps (\tilde{f}, f) , with $\tilde{f}: E_1 \rightarrow E_2$ and $f: B_1 \rightarrow B_2$, such that

(1) the following diagram commutes

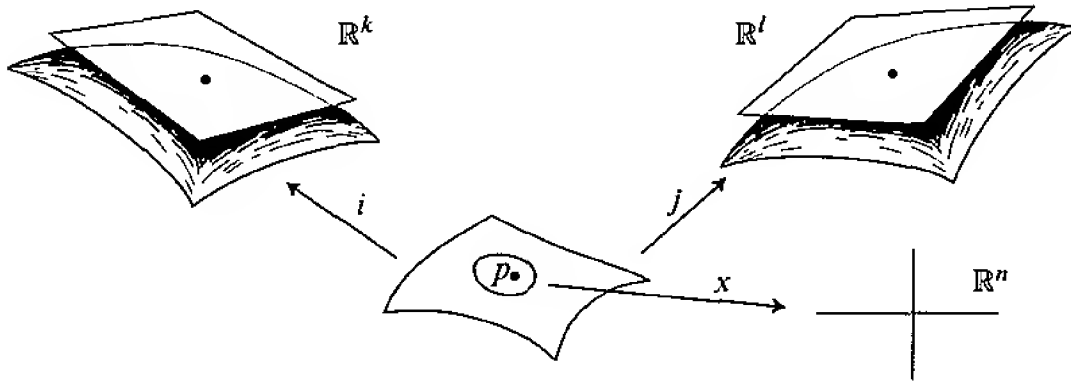
$$\begin{array}{ccc} E_1 & \xrightarrow{\tilde{f}} & E_2 \\ \pi_1 \downarrow & & \downarrow \pi_2 \\ B_1 & \xrightarrow{f} & B_2 \end{array},$$

(2) $\tilde{f}: \pi_1^{-1}(p) \rightarrow \pi_2^{-1}(f(p))$ is a linear map.

The pair (f_*, f) is a bundle map from $T\mathbb{R}^k$ to $T\mathbb{R}^l$ for any differentiable $f: \mathbb{R}^k \rightarrow \mathbb{R}^l$. If $M^n \subset \mathbb{R}^k$ and $N^m \subset \mathbb{R}^l$ are submanifolds, $i: M \rightarrow \mathbb{R}^k$ and $j: N \rightarrow \mathbb{R}^l$ are the inclusions, and the map f satisfies $f(M) \subset N$, then f_*

*There are actually several possible choices, depending on whether one is considering all bundles at once, fixed bundles over various spaces, or a fixed base space with varying bundles. Thus f may be restricted to be an isomorphism on fibres and f to be the identity, or a homeomorphism. The relations between some of these cases are considered in the problems.

takes $T(M, i)$ to $T(N, j)$; to see this, just remember that $v \in T(M, i)_p$ is the tangent vector of a curve c in M , so $f_*(v)$ is the tangent vector of the curve $f \circ c$ in N , and consequently $f_*(v) \in T(N, j)$. In this way we obtain a bundle map from $T(M, i)$ to $T(N, j)$. Actually, it would have sufficed to begin with a C^∞ function $f: M \rightarrow N$, since f can be extended to \mathbb{R}^k locally. In fact, this construction could be generalized much further, to the case where i and j are merely imbeddings of two abstract manifolds M and N , and $f: M \rightarrow N$ is C^∞ ; we just consider the function $j \circ f \circ i^{-1}: i(M) \rightarrow i(N)$ and extend it locally to \mathbb{R}^k . The case which we want to examine most carefully is the simplest: where $M = N$ and f is the identity, while i and j are two imbeddings of M in \mathbb{R}^k and \mathbb{R}^l , respectively. Elements of $T(M, i)_p$ are of the form $(i \circ x^{-1})_*(w)$



for $w \in \mathbb{R}^n_{x(p)}$, while elements of $T(M, j)_p$ are of the form $(j \circ x^{-1})_*(w)$ for $w \in \mathbb{R}^n_{x(p)}$. If we map

$$(i \circ x^{-1})_*(w) \mapsto (j \circ x^{-1})_*(w)$$

we obtain a bundle map from $T(M, i)|U$ to $T(M, j)|U$, which is obviously an equivalence. The map $(M, i)_p \rightarrow (M, j)_p$ induced on fibres is independent of the coordinate system x , for if (y, V) is another coordinate system, then

$$\begin{array}{ccc} (i \circ y^{-1})_*(w) & = & (i \circ x^{-1})_*((x \circ y^{-1})_*(w)) \\ \downarrow & & \downarrow \\ (j \circ y^{-1})_*(w) & = & (j \circ x^{-1})_*((x \circ y^{-1})_*(w)). \end{array}$$

We can therefore put all these maps together, and obtain an *equivalence* from $T(M, i)$ to $T(M, j)$. In other words, the dependence of $T(M, i)$ on i is almost illusory; we could abbreviate $T(M, i)$ to TM , if we agreed that TM really denotes an equivalence class of bundles, rather than one bundle. That is the

sort of thing an algebraist might do, and it is undoubtedly ugly. What we would like to do is to get a single bundle for each M , in some natural way, which has all the properties any one of these particular bundles $T(M, i)$ has. Can we do this? Yes, we can. When we do, $T\mathbb{R}^n$ will be different from our old definition (namely, $\varepsilon^n(\mathbb{R}^n)$), and so will f_* for $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, so in stating our result precisely we will write “old f_* ” when necessary.

1. THEOREM. It is possible to assign to each n -manifold M an n -plane bundle TM over M , and to each C^∞ map $f: M \rightarrow N$ a bundle map (f_*, f) , such that:

- (1) If $1: M \rightarrow M$ is the identity, then $1_*: TM \rightarrow TM$ is the identity. If $g: N \rightarrow P$, then $(g \circ f)_* = g_* \circ f_*$.
- (2) There are equivalences $t^n: T\mathbb{R}^n \rightarrow \varepsilon^n(\mathbb{R}^n)$ such that for every C^∞ function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ the following commutes.

$$\begin{array}{ccc} T\mathbb{R}^n & \xrightarrow{f_*} & T\mathbb{R}^m \\ t^n \downarrow & & \downarrow t^m \\ \varepsilon^n(\mathbb{R}^n) & \xrightarrow{\text{old } f_*} & \varepsilon^m(\mathbb{R}^m) \end{array}$$

- (3) If $U \subset M$ is an open submanifold, then TU is equivalent to $(TM)|U$, and for $f: M \rightarrow N$ the map $(f|U)_*: TU \rightarrow TN$ is just the restriction of f_* . More precisely, there is an equivalence $TU \simeq (TM)|U$ such that the following diagrams commute, where $i: U \rightarrow M$ is the inclusion.*

$$\begin{array}{ccc} TU & \xrightarrow{i_*} & TM \\ \simeq \searrow & & \nearrow \subset \\ & (TM)|U & \end{array} \qquad \begin{array}{ccc} TU & \xrightarrow{(f|U)_*} & TN \\ i_* \searrow & & \nearrow f_* \\ & TM & \end{array}$$

PROOF. The construction of TM is an ingenious, though quite natural, subterfuge. We will obtain a single bundle for TM , but the *elements* of TM will each be large equivalence classes.

* When using the notation f_* , it must be understood that the symbol “ f ” really refers to a triple (f, M, N) where $f: M \rightarrow N$. The identity map 1 of U to itself and the inclusion map $i: U \rightarrow M$ have to be considered as different, since the maps $1_*: TU \rightarrow TU$ and $i_*: TU \rightarrow TM$ are certainly different (they map TU into two different sets).

The construction is much easier to understand if we first imagine that we *already had* our bundles TM . Then if (x, U) is a coordinate system, we would have a map $x_*: TU \rightarrow T(x(U))$, and this would be an equivalence (with inverse $(x^{-1})_*$). Since TU should be essentially $(TM)|_U$, and $T(x(U))$ should essentially be $x(U) \times \mathbb{R}^n$, a point $e \in \pi^{-1}(p)$ would be taken by x_* to some $(x(p), v)$. Here v is just an element of \mathbb{R}^n (and every v would occur, since x_* maps $\pi^{-1}(p)$ isomorphically onto $\{p\} \times \mathbb{R}^n$). If y is another coordinate system, then $y_*(e)$ would be $(y(p), w)$ for some $w \in \mathbb{R}^n$. We can easily figure out what the relationship between v and w would be; since $(x(p), v)$ is taken to $(y(p), w)$ by $y_* \circ x_*^{-1} = (y \circ x^{-1})_*$, and $(y \circ x^{-1})_*$ is supposed to be the old $(y \circ x^{-1})_*$, we would have

$$(a) \quad w = D(y \circ x^{-1})(x(p))(v).$$

This condition makes perfect sense without any mention of bundles. It is the clue which enables us to now define TM .

If x and y are coordinate systems whose domains contain p , and $v, w \in \mathbb{R}^n$, we *define*

$$(x, v) \sim_p (y, w) \quad \text{if (a) is satisfied.}$$

It is easy to check (using the chain rule) that \sim_p is an equivalence relation; the equivalence class of (x, v) will be denoted by $[x, v]_p$. These equivalence classes will be called tangent vectors at p , and TM is defined to be the set of all tangent vectors at all points $p \in M$; the map π takes \sim_p equivalence classes to p . We define a vector space structure on $\pi^{-1}(p)$ by the formulas

$$\begin{aligned} [x, v]_p + [x, w]_p &= [x, v + w]_p \\ a \cdot [x, v]_p &= [x, a \cdot v]_p; \end{aligned}$$

this definition is independent of the particular coordinate system x or y , because $D(y \circ x^{-1})(x(p))$ is an isomorphism from \mathbb{R}^n to \mathbb{R}^n .

Our definition of TM provides a one-one onto map

$$(b) \quad t_x: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n, \quad \text{namely} \quad [x, v]_q \mapsto (q, v).$$

We want this to be a homeomorphism, so we want $t_x^{-1}(A)$ to be open for every open $A \subset U \times \mathbb{R}^n$, and thus we want any union of such sets to be open. There is a metric with exactly these sets as open sets, but it is a little ticklish to produce, so we leave this one part of the proof to Problem 1.

We now have a bundle $\pi: TM \rightarrow M$. We will denote the fibre $\pi^{-1}(p)$ by M_p , in conformity with the notation \mathbb{R}^n_p , though TM_p might be better. If

$f: M \rightarrow N$, and (x, U) and (y, V) are coordinate systems around p and $f(p)$, respectively, we define

$$(c) \quad f_*([x, v]_p) = [y, D(y \circ f \circ x^{-1})(x(p))(v)]_{f(p)}.$$

Of course, it must be checked that this definition is independent of x and y (the chain rule again).

Condition (1) of our theorem is obvious.

To prove (2), we define t^n to be t_I , where I is the identity map of \mathbb{R}^n and t_x is defined in (b); it is trivial, though perhaps confusing to the novice, to prove commutativity of the diagram.

Condition (3) is practically obvious also. In fact, the fibre of TU over $p \in U$ is almost exactly the same as the fibre of TM over p ; the only difference is that each equivalence class for M contains some extra members, since in M there are more coordinate systems around p than there are in $U \subset M$. ♦

Henceforth, the bundle $\pi: TM \rightarrow M$ will be called the **tangent bundle** of M . If $i: M \rightarrow \mathbb{R}^k$ is an imbedding, then TM is equivalent to $T(M, i)$. In fact, if (x, U) is a coordinate system around p , and I is the identity coordinate system of \mathbb{R}^k , then

$$\begin{aligned} i_*([x, v]_p) &= [I, D(i \circ x^{-1})(x(p))(v)]_{i(p)} && \text{by (c)} \\ t^n &= t_I \quad \downarrow \\ (i(p), D(i \circ x^{-1})(x(p))(v)) &\in (M, i)_p; \end{aligned}$$

the composition $t^n i_*$ is easily seen to be an equivalence. But $T(M, i)$ will play no further role in this story—the abstract substitute TM will always be used instead.

Having succeeded in producing a bundle over each M , which is equivalent to $T(M, i)$, we next ask how fortuitous this was. Can one find other bundles with the same properties? The answer is yes, and we proceed to define two different such bundles.

For the first example, we consider curves $c: (-\varepsilon, \varepsilon) \rightarrow M$, each defined on some interval around 0, with $c(0) = p$. If (x, U) is a coordinate system around p , we define

$$c_1 \underset{p}{\approx} c_2 \quad \text{if and only if:} \quad \begin{array}{l} x \circ c_1 \text{ and } x \circ c_2, \text{ mapping } \mathbb{R} \text{ to } \mathbb{R}^n, \\ \text{have the same derivative at 0.} \end{array}$$

The equivalence classes, for all $p \in M$, will be the elements of our new bundle, $T'M$. For $f: M \rightarrow N$ there is a map $f_\#$ taking the $\underset{p}{\approx}$ equivalence class

of c to the $\widetilde{f(p)}$ equivalence class of $f \circ c$. Without bothering to check details, we can already see that this example is “really the same” as TM —

$[x, v]_p$ corresponds to: the \approx_p equivalence class of $x^{-1} \circ \gamma$,
where γ is a curve in \mathbb{R}^n with $\gamma'(0) = v$;

under this correspondence, $f_\#$ corresponds to f_* .

In the second example, things are not so simple. We define a tangent vector at p to be a linear operator ℓ which operates on all C^∞ functions f and which is a “derivation at p ”:

$$\ell(fg) = f(p)\ell(g) + g(p)\ell(f).$$

We have already seen that the operators $\ell = \partial/\partial x^i|_p$ have this property. For these operators, clearly $\ell(f) = \ell(g)$ if $f = g$ in a neighborhood of p . This condition is actually true for any derivation ℓ . For, suppose that $f = 0$ in a neighborhood of p . There is a C^∞ function $h: M \rightarrow \mathbb{R}$ with $h(p) = 1$ and support $h \subset f^{-1}(0)$. Then

$$0 = \ell(0) = \ell(fh) = f(0)\ell(h) + h(0)\ell(f) = 0 + \ell(f).$$

Thus, if $f = g$ in a neighborhood of 0, then $0 = \ell(f - g) = \ell(f) - \ell(g)$. If f is defined only in a neighborhood of p , we may use this trick to define $\ell(f)$: choose h to be 1 on a neighborhood of p , with support $h \subset f^{-1}(0)$, and define $\ell(f)$ as $\ell(fh)$.

The set of all such operators is a vector space, but it is not *a priori* clear what its dimension is. This comes out of the following.

2. LEMMA. Let f be a C^∞ function in a convex open neighborhood U of 0 in \mathbb{R}^n , with $f(0) = 0$. Then there are C^∞ functions $g_i: U \rightarrow \mathbb{R}$ with

$$(1) \quad f(x^1, \dots, x^n) = \sum_{i=1}^n x^i g_i(x^1, \dots, x^n) \quad \text{for } x \in U,$$

$$(2) \quad g_i(0) = D_i f(0).$$

(The second condition actually follows from the first.)

PROOF. For $x \in U$, let $h_x(t) = f(tx)$; this is defined for $0 \leq t \leq 1$, since U is convex. Then

$$f(x) = f(x) - f(0) = \int_0^1 h_x'(t) dt = \int_0^1 \sum_{i=1}^n D_i f(tx) \cdot x^i dt.$$

Therefore we can let $g(x) = \int_0^1 D_i f(tx) dt$. ♦

3. THEOREM. The set of all linear derivations at $p \in M^n$ is an n -dimensional vector space. In fact, if (x, U) is a coordinate system around p , then

$$\left. \frac{\partial}{\partial x^1} \right|_p, \dots, \left. \frac{\partial}{\partial x^n} \right|_p$$

span this vector space, and any derivation ℓ can be written

$$\ell = \sum_{i=1}^n \ell(x^i) \cdot \left. \frac{\partial}{\partial x^i} \right|_p,$$

(so ℓ is determined by the numbers $\ell(x^i)$).

PROOF. Notice that

$$\ell(1) = \ell(1 \cdot 1) = 1 \cdot \ell(1) + 1 \cdot \ell(1),$$

so $\ell(1) = 0$. Hence $\ell(c) = c \cdot \ell(1) = 0$ for any constant function c on U . Consider the case where $M = \mathbb{R}^n$ and $p = 0$. Assume U is convex. Given f on U , choose g_i as in Lemma 2, for the function $f - f(0)$. Then

$$\begin{aligned} \ell(f) &= \ell(f - f(0)) = \ell\left(\sum_{i=1}^n I^i g_i\right) \quad (I^i \text{ denotes the } i^{\text{th}} \\ &\quad \text{coordinate function}) \\ &= \sum_{i=1}^n [\ell(I^i)g_i(0) + I^i(0)\ell(g_i)] \\ &= \sum_{i=1}^n \ell(I^i) \frac{\partial f}{\partial I^i}(0) + 0. \end{aligned}$$

This shows that $\partial/\partial I^i|_0$ span the vector space; they are clearly linearly independent. It is a simple exercise to use the coordinate system x to transfer this result from \mathbb{R}^n to M . ♦

From Theorem 3 we can see that, once again, a bundle constructed from all derivations at all points of M is “really the same” as TM . We can let

$$\ell = \sum_{i=1}^n a^i \left. \frac{\partial}{\partial x^i} \right|_p \quad \text{correspond to} \quad [x, a]_p;$$

the formula

$$\left. \frac{\partial}{\partial x^i} \right|_p = \sum_{j=1}^n \frac{\partial y^j}{\partial x^i}(p) \left. \frac{\partial}{\partial y^j} \right|_p,$$

derived in Chapter 2, shows that

$$\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p = \sum_{i=1}^n b^i \frac{\partial}{\partial y^i} \Big|_p \quad \text{if and only if} \quad b^j = \sum_{i=1}^n a^i \frac{\partial y^j}{\partial x^i}(p),$$

and this is precisely the equation which says that $(x, a) \underset{p}{\sim} (y, b)$. It is easily checked that under this correspondence, the map which corresponds to f_* can be defined as follows:

$$[f_*(\ell)](g) = \ell(g \circ f).$$

Notice that if x denotes the identity coordinate system on \mathbb{R}^n , then $\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p$ corresponds to a_p when we identify $T\mathbb{R}^n$ with $\varepsilon^n(\mathbb{R}^n)$.

We will usually make no distinction whatsoever between a tangent vector $v \in M_p$ and the linear derivation it corresponds to, that is, between $[x, a]_p$ and

$$\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p;$$

consequently, we will not hesitate to write $v(f)$ for a differentiable function f defined in a neighborhood of p . In fact, a tangent vector is often most easily described by telling what derivation it corresponds to, and the map f_* is often most easily analyzed from the relation

$$(f_*v)(g) = v(g \circ f).$$

It is customary to denote the identity coordinate system on \mathbb{R}^1 by t , and to write

$$\frac{d}{dt} \Big|_{t_0} \quad \text{for} \quad \frac{\partial}{\partial t} \Big|_{t_0};$$

this is a basis for \mathbb{R}_{t_0} . If $c: \mathbb{R} \rightarrow M$ is a differentiable curve, then

$$c_* \left(\frac{d}{dt} \Big|_{t_0} \right) \in M_{c(t_0)}$$

is called the *tangent vector to c at t_0* . We will denote it by the suggestive symbol

$$\frac{dc}{dt} \Big|_{t_0}.$$

This symbol will be subjected to the standard abuses one finds (unexplained) in calculus textbooks: the symbol

$$\frac{dc}{dt} \quad \text{will often stand for} \quad \left. \frac{dc}{dt} \right|_t,$$

the subscript “ t ” now denoting a particular number $t \in \mathbb{R}$, as well as the identity coordinate system.

As you might well expect, it is no accident that our second and third examples turned out to be “really the same” as TM . There is a general theorem that all “reasonable” examples will have this property, but it is a little delicate to state, and quite a mess to prove, so it has been quarantined in an Addendum to this chapter.

The tangent bundle TM of a C^∞ manifold has a little more structure than an arbitrary n -plane bundle. Since TM locally looks like $U \times \mathbb{R}^n$, clearly TM is itself a manifold; there is, moreover, a natural way to put a C^∞ structure on TM . If $x: U \rightarrow \mathbb{R}^n$ is a chart on M , then every element $v \in (TM)|_U$ is uniquely of the form

$$v = \sum_{i=1}^n a^i \left. \frac{\partial}{\partial x^i} \right|_p, \quad p = \pi(v).$$

Let us denote a^i by $\dot{x}^i(v)$. Then the map

$$v \mapsto (x^1(\pi(v)), \dots, x^n(\pi(v)), \dot{x}^1(v), \dots, \dot{x}^n(v)) \in \mathbb{R}^{2n}$$

is a homeomorphism from $(TM)|_U$ to $x(U) \times \mathbb{R}^n$. This map, $(x \circ \pi, \dot{x})$, is simply the map x_* when we identify TU with $U \times \mathbb{R}^n$ in the standard way. If (y, V) is another coordinate system, and

$$v = \sum_{j=1}^n b^j \left. \frac{\partial}{\partial y^j} \right|_p,$$

then, as we have already seen,

$$b^j = \sum_{i=1}^n a^i \frac{\partial y^j}{\partial x^i}(p) = \sum_{i=1}^n a^i D_i(y^j \circ x^{-1})(x(p)).$$

This shows that if $(t, a) = (t^1, \dots, t^n, a^1, \dots, a^n) \in \mathbb{R}^{2n}$, then

$$\begin{aligned} y_* \circ (x_*)^{-1}(t, a) \\ = (y \circ x^{-1}(t), \sum_{i=1}^n a^i D_i(y^1 \circ x^{-1})(t), \dots, \sum_{i=1}^n a^i D_i(y^n \circ x^{-1})(t)). \end{aligned}$$

This expression shows that $y_* \circ (x_*)^{-1}$ is C^∞ .

We thus have a collection of C^∞ -related charts on TM , which can be extended to a maximal atlas.

With this C^∞ structure, the local trivializations x_* are C^∞ . In general, a vector bundle $\pi: E \rightarrow B$ is called a C^∞ vector bundle if E and B are C^∞ manifolds and there are C^∞ local trivializations in a neighborhood of each point. It follows that $\pi: E \rightarrow B$ is C^∞ .

Recall that a section of a bundle $\pi: E \rightarrow B$ is a continuous function $s: B \rightarrow E$ such that $\pi \circ s = \text{identity of } B$; for C^∞ vector bundles we can also speak of C^∞ sections. A section of TM is called a vector field on M ; for submanifolds M of \mathbb{R}^n , a vector field may be pictured as a continuous selection of arrows tangent to M . The theorem that you can't comb the hair on a sphere just states that



there is no vector field on S^2 which is everywhere non-zero. We have shown that there do not exist two vector fields on the Möbius strip which are everywhere linearly independent.

Vector fields are customarily denoted by symbols like X , Y , or Z , and the vector $X(p)$ is often denoted by X_p (sometimes X may be used to denote a single vector, in some M_p). If we think of TM as the set of derivations, then for any coordinate system (x, U) , we have

$$X(p) = \sum_{i=1}^n a^i(p) \left. \frac{\partial}{\partial x^i} \right|_p \quad \text{for all } p \in U.$$

The functions a^i are continuous or C^∞ if and only if $X: U \rightarrow TM$ is continuous or C^∞ .

If X and Y are two vector fields, we define a new vector field $X + Y$ by

$$(X + Y)(p) = X(p) + Y(p).$$

Similarly, if $f: M \rightarrow \mathbb{R}$, we define the vector field fX by

$$(fX)(p) = f(p)X(p).$$

Clearly $X + Y$ and fX are C^∞ if X , Y , and f are C^∞ . On U we can write

$$X = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i},$$

the symbol $\partial/\partial x^i$ now denoting the vector field

$$p \mapsto \left. \frac{\partial}{\partial x^i} \right|_p.$$

If $f: M \rightarrow \mathbb{R}$ is a C^∞ function, and X is a vector field, then we can define a new function $\bar{X}(f): M \rightarrow \mathbb{R}$ by letting X operate on f at each point:

$$\bar{X}(f)(p) = X_p(f).$$

It is not hard to check that if X is a C^∞ vector field, then $\bar{X}(f)$ is C^∞ for every C^∞ function f ; indeed, if locally

$$X(p) = \sum_{i=1}^n a^i(p) \left. \frac{\partial}{\partial x^i} \right|_p,$$

then

$$\bar{X}(f) = \sum_{i=1}^n a^i \frac{\partial f}{\partial x^i},$$

which is a sum of products of C^∞ functions. Conversely, if $\bar{X}(f)$ is C^∞ for every C^∞ function f , then X is a C^∞ vector field (since $\bar{X}(x^i) = a^i$).

Let \mathcal{F} denote the set of all C^∞ functions on M . We have just seen that a C^∞ vector field X gives rise to a function $\bar{X}: \mathcal{F} \rightarrow \mathcal{F}$. Clearly,

$$\begin{aligned} \bar{X}(f + g) &= \bar{X}(f) + \bar{X}(g) \\ \bar{X}(fg) &= f\bar{X}(g) + g\bar{X}(f); \end{aligned}$$

thus \bar{X} is a "derivation" of the ring \mathcal{F} . Often, a C^∞ vector field X is identified with the derivation \bar{X} . The reason for this is that if $A: \mathcal{F} \rightarrow \mathcal{F}$ is any derivation, then $A = \bar{X}$ for a unique C^∞ vector field X . In fact, we clearly must define

$$X_p(f) = A(f)(p),$$

and the operator X_p thus defined is a derivation at p .

The tangent bundle is the true beginning of the study of differentiable manifolds, and you should not read further until you grok it.* The next few chapters constitute a detailed study of this bundle. One basic theme in all these chapters is that any structure one can put on a vector space leads to a structure on any vector bundle, in particular on the tangent bundle of a manifold. For the present, we will discuss just one new concept about manifolds, which arises in this very way from the notion of “orientation” in a vector space.

The non-singular linear maps $f: V \rightarrow V$ from a finite dimensional vector space to itself fall into two groups, those with $\det f > 0$, and those with $\det f < 0$; linear transformations in the first group are called **orientation preserving** and the others are called **orientation reversing**. A simple example of the latter is the map $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by $f(x) = (x^1, \dots, x^{n-1}, -x^n)$ (reflection in the hyperplane $x^n = 0$). There is no way to pass continuously between these two groups: if we identify linear maps $\mathbb{R}^n \rightarrow \mathbb{R}^n$ with $n \times n$ matrices, and thus with \mathbb{R}^{n^2} , then the orientation preserving and orientation reversing maps are disjoint open subsets of the set of all non-singular maps (those with $\det \neq 0$). The terminology “orientation preserving” is a bit strange, since we have not yet defined anything called “orientation”, which is being preserved. The problem becomes more acute if we want to define orientation preserving isomorphisms between two different (but isomorphic) vector spaces V and W ; this clearly makes no sense unless we supply V and W with more structure.

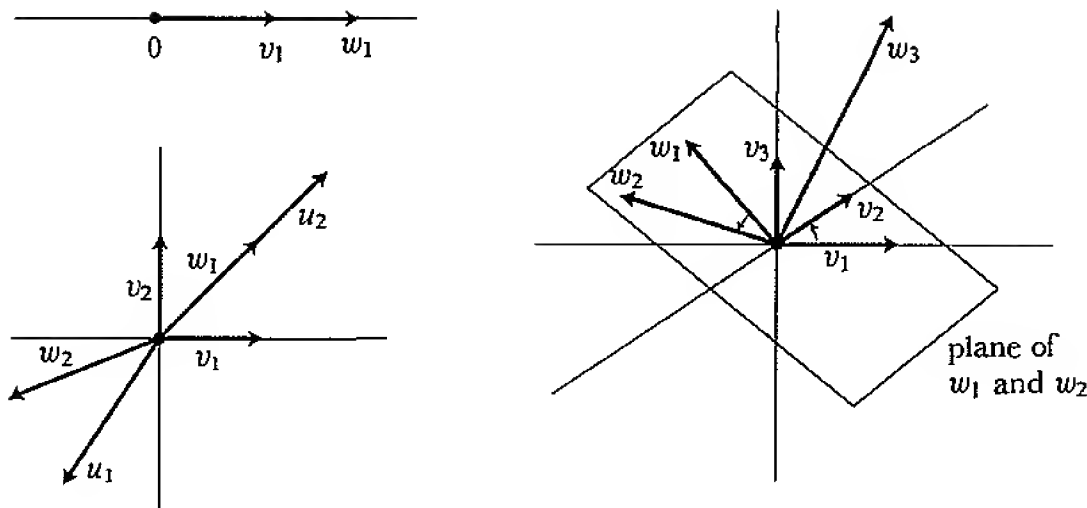
To provide this extra structure, we note that two ordered bases (v_1, \dots, v_n) and (v'_1, \dots, v'_n) for V determine an isomorphism $f: V \rightarrow V$ with $f(v_i) = v'_i$; the matrix $A = (a_{ij})$ of f is given by the equations

$$v'_i = \sum_{j=1}^n a_{ji} v_j.$$

We call (v_1, \dots, v_n) and (v'_1, \dots, v'_n) **equally oriented** if $\det A > 0$ (i.e., if f is orientation preserving) and **oppositely oriented** if $\det A < 0$.

The relation of being equally oriented is clearly an equivalence relation, dividing the collection of all ordered bases into just two equivalence classes. Either of these two equivalence classes is called an **orientation** for V . The class to which (v_1, \dots, v_n) belongs will be denoted by $[v_1, \dots, v_n]$, so that if μ is an orientation of V , then $(v_1, \dots, v_n) \in \mu$ if and only if $[v_1, \dots, v_n] = \mu$. If μ denotes one

* A cult word of the sixties, “grok” was coined, purportedly as a word from the Martian language, by Robert A. Heinlein in his pop science fiction novel *Stranger in a Strange Land*. Its sense is nicely conveyed by the definition in *The American Heritage Dictionary*: “To understand profoundly through intuition or empathy”.



Examples of equally oriented ordered bases in \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 .

orientation of V , the other will be denoted by $-\mu$, and the orientation $[e_1, \dots, e_n]$ for \mathbb{R}^n will be called the “standard orientation”.

Now if (V, μ) and (W, ν) are two n -dimensional vector spaces, together with orientations, an isomorphism $f: V \rightarrow W$ is called **orientation preserving** (with respect to μ and ν) if $[f(v_1), \dots, f(v_n)] = \nu$ whenever $[v_1, \dots, v_n] = \mu$; if this holds for any one (v_1, \dots, v_n) , it clearly holds for all.

For the trivial bundle $\varepsilon^n(X) = X \times \mathbb{R}^n$ we can put the “standard orientation” $[(x, e_1), \dots, (x, e_n)]$ on each fibre $\{x\} \times \mathbb{R}^n$. If $f: \varepsilon^n(X) \rightarrow \varepsilon^n(X)$ is an equivalence, and X is connected, then f is either orientation preserving or orientation reversing on each fibre, for if we define the functions $a_{ij}: X \rightarrow \mathbb{R}$ by

$$f(x, e_i) = \sum_{j=1}^n a_{ji}(x) \cdot (x, e_j),$$

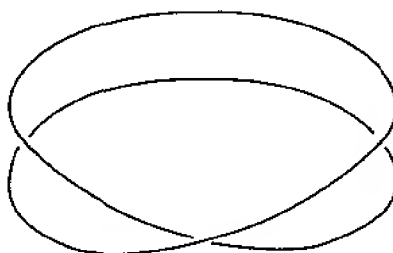
then $\det(a_{ij}): X \rightarrow \mathbb{R}$ is continuous and never 0. If $\pi: E \rightarrow B$ is a non-trivial n -plane bundle, an **orientation** μ of E is defined to be a collection of orientations μ_p for $\pi^{-1}(p)$ which satisfy the following “compatibility condition” for any open connected set $U \subset B$:

If $t: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$ is an equivalence, and the fibres of $U \times \mathbb{R}^n$ are given the standard orientation, then t is either orientation preserving or orientation reversing on all fibres.

Notice that if this condition is satisfied for a certain t , and $t': \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$ is another equivalence, then t' automatically satisfies the same condition, since

$t' \circ t^{-1}: U \times \mathbb{R}^n \rightarrow U \times \mathbb{R}^n$ is an equivalence. This shows that the orientations μ_p define an orientation of E if the compatibility condition holds for a collection of sets U which cover B .

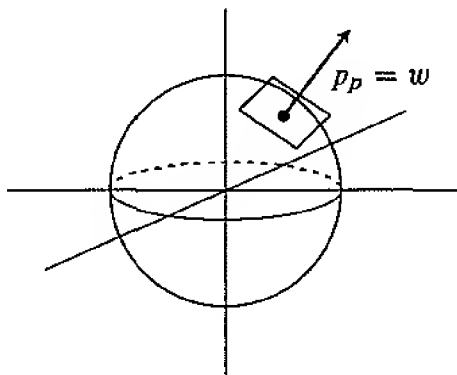
If a bundle E has orientation $\mu = \{\mu_p\}$, it has another orientation $-\mu = \{-\mu_p\}$, but not every bundle has an orientation. For example, the Möbius strip, considered as a 1-dimensional bundle over S^1 , has no orientation. For, although the Möbius strip has no non-zero section, we can pick two vectors from each fibre so that the totality A looks like two sections. For example, we can let A be $[0, 1] \times \{-1, 1\}$ with $(0, a)$ identified with $(1, -a)$; then A just looks like the boundary of the Möbius strip obtained from $[0, 1] \times [-1, 1]$. If



we had compatible orientations μ_p , we could define a section $s: S^1 \rightarrow M$ by choosing $s(p)$ to be the unique vector $s(p) \in A \cap \pi^{-1}(p)$ with $[s(p)] = \mu_p$.

A bundle is called **orientable** if it has an orientation, and **non-orientable** otherwise; an **oriented bundle** is just a pair (ξ, μ) where μ is an orientation for ξ . This definition can be applied, in particular, to the tangent bundle TM of a C^∞ manifold M . In this case, we call M itself **orientable** or **non-orientable** depending on whether TM is orientable or non-orientable; an orientation of TM is also called an **orientation** of M , and an **oriented manifold** is a pair (M, μ) where μ is an orientation for TM .

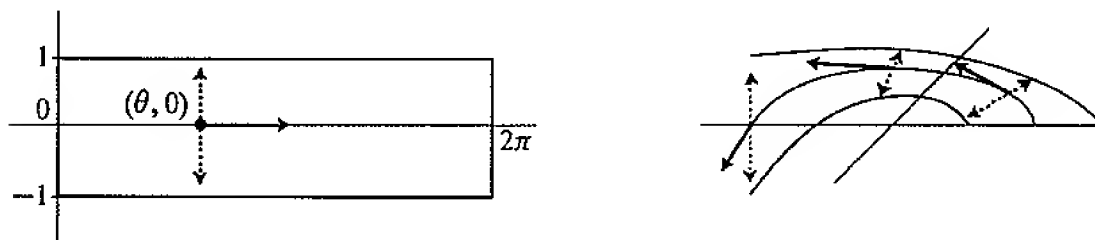
The manifold \mathbb{R}^n is orientable, since $T\mathbb{R}^n \simeq \varepsilon^n(\mathbb{R}^n)$, on which we have the standard orientation. The sphere $S^{n-1} \subset \mathbb{R}^n$ is also orientable. To see this we



note that for each $p \in S^{n-1}$ the vector $w = p_p \in \varepsilon^n(\mathbb{R}^n) \simeq T\mathbb{R}^n$ is *not* in $i_*(S^{n-1}_p) \in T\mathbb{R}^n_p$ (Problem 21), so for $v_1, \dots, v_{n-1} \in S^{n-1}_p$ we can define $(v_1, \dots, v_{n-1}) \in \mu_p$ if and only if $(w, i_*(v_1), \dots, i_*(v_{n-1}))$ is in the standard orientation of \mathbb{R}^n_p . The orientation $\mu = \{\mu_p : p \in S^{n-1}\}$ thus defined is called the “standard orientation” of S^{n-1} .

The torus $S^1 \times S^1$ is another example of an orientable manifold. This can be seen by noting that for any two manifolds M_1 and M_2 the fibre $(M_1 \times M_2)_p$ of $T(M_1 \times M_2)$ can be written as $V_{1p} \oplus V_{2p}$ where $(\pi_i)_*: V_{ip} \rightarrow (M_i)_p$ is an isomorphism and the subspaces V_{ip} vary continuously (Problem 26). Since TS^1 is trivial, this shows that $T(S^1 \times S^1)$ is also trivial, and consequently orientable. Any n -holed torus is also orientable—the proof is presented in Problem 16, which also discusses the tangent bundle of a manifold-with-boundary.

The Möbius strip M is the simplest example of a non-orientable 2-manifold. For the imbedding of M considered previously we have already seen that on the



subset $S = \{(2 \cos \theta, 2 \sin \theta, 0)\} \subset M$ there are continuously varying vectors v_p , but that it is impossible to choose continuously from among the dashed vectors $w_p = f_*(0, 1)_{(\theta, 0)}$ and their negatives. If we had orientations μ_p for $p \in S$, then we could simply choose w_p if $[v_p, w_p] = \mu_p$ and $-w_p$ otherwise.

The projective plane \mathbb{P}^2 must be non-orientable also, since it contains the Möbius strip (for any orientable bundle $\xi = \pi: E \rightarrow B$, the restriction $\xi|_{B'}$ to any subset $B' \subset B$ is also orientable). Non-orientability of \mathbb{P}^2 can be seen in another way, by considering the “antipodal map” $A: S^2 \rightarrow S^2$ defined by $A(p) = -p$. This map is just the restriction of a linear map $\bar{A}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ defined by the same formula. The map $A_*: S^2_p \rightarrow S^2_{A(p)}$ is just $(p, v) \mapsto (\bar{A}(p), \bar{A}(v))$, when S^2_p is identified with a subspace of $\{p\} \times \mathbb{R}^3$. The map \bar{A} is orientation reversing, so if $v_i = (p, u_i) \in S^2_p$, the bases

$$(u_1, u_2, p) \quad \text{and} \quad (\bar{A}(u_1), \bar{A}(u_2), \bar{A}(p))$$

are oppositely oriented. This shows that if μ is the standard orientation of S^2 and $[v_1, v_2] \in \mu_p$, then $[A_*v_1, A_*v_2] \in -\mu_{A(p)}$. Thus the map $A: S^2 \rightarrow S^2$ is

“orientation reversing” (the notion of an **orientation preserving** or **orientation reversing** map $f: M \rightarrow N$ makes sense for any imbedding f of one oriented manifold into another oriented manifold of the same dimension). From this fact it follows easily that \mathbb{P}^2 is not orientable: If \mathbb{P}^2 had an orientation $\nu = \{\nu_{[p]}\}$ and $g: S^2 \rightarrow \mathbb{P}^2$ is the map $p \mapsto [p]$, then we could define an orientation $\{\bar{\mu}_p\}$ on S^2 by requiring g to be orientation preserving; the map A would then be orientation preserving with respect to $\bar{\mu}$, which is impossible, since $\bar{\mu} = \mu$ or $-\mu$.

For projective 3-space \mathbb{P}^3 the situation is just the opposite. In this case, the antipodal map $A: S^3 \rightarrow S^3$ is orientation preserving. If $g: S^3 \rightarrow \mathbb{P}^3$ is the map $p \mapsto [p]$, we obviously can define orientations ν_p for \mathbb{P}^3 by requiring g to be orientation preserving. In general, these same arguments show that \mathbb{P}^n is orientable for n odd and non-orientable for n even.

There is a more “elementary” definition of orientability, which does not use the tangent bundle of M at all. According to this definition, M is orientable if there is a subset \mathcal{A}' of the atlas \mathcal{A} for M such that

- (1) the domains of all $(x, U) \in \mathcal{A}'$ cover M ,
- (2) for all (x, U) and $(y, V) \in \mathcal{A}'$,

$$\det \left(\frac{\partial y^i}{\partial x^j} \right) > 0 \quad \text{on} \quad U \cap V.$$

An orientation μ of TM allows us to distinguish the subset \mathcal{A}' as the collection of all (x, U) for which $x_*: TM|U \rightarrow T(x(U)) \simeq x(U) \times \mathbb{R}^n$ is orientation preserving (when $x(U) \times \mathbb{R}^n$ is given the standard orientation). Condition (2) holds, because it is just the condition that $(y \circ x^{-1})_*: T(x(U)) \rightarrow T(y(V))$ is orientation preserving. Conversely, given \mathcal{A}' we can orient the fibres of $TM|U$ in such a way that x_* is orientation preserving, and obtain an orientation of TM . Although our original definition is easier to picture geometrically, the determinant condition will be very important later on.

ADDENDUM

EQUIVALENCE OF TANGENT BUNDLES

The fact that all reasonable candidates for the tangent bundle of M turn out to be essentially the same is stated precisely as follows.

4. THEOREM*. If we have a bundle $T'M$ over M for each M , and a bundle map (f_{\sharp}, f) for each C^{∞} map $f: M \rightarrow N$ satisfying

- (1) of Theorem 1,
- (2) of Theorem 1, for certain equivalences t'^n ,
- (3) of Theorem 1, for certain equivalences $T'U \simeq (T'M)|U$,

then there are equivalences

$$e_M: TM \rightarrow T'M$$

such that the following diagram commutes for every C^{∞} map $f: M \rightarrow N$.

$$\begin{array}{ccc} TM & \xrightarrow{f_*} & TN \\ e_M \downarrow & & \downarrow e_N \\ T'M & \xrightarrow{f_{\sharp}} & T'N \end{array}$$

PROOF. The details of this proof are so horrible that you should probably skip it (and you should definitely quit when you get bogged down); the welcome symbol \diamond occurs quite a ways on. Nevertheless, the idea behind the proof is simple enough. If (x, U) is a chart on M , then both $(TM)|U$ and $(T'M)|U$ “look like” $x(U) \times \mathbb{R}^n$, so there ought to be a map taking the fibres of one to the fibres of the other. What we have to hope is that our conditions on TM and $T'M$ make them “look alike” in a sufficiently strong way for this idea to really work out. Those who have been through this sort of rigamarole before know (i.e., have faith) that it’s going to work out; those for whom this sort of proof is a new experience should find it painful and instructive.

Functorites will notice that Theorems 1 and 4 say that there is, up to natural equivalence, a unique functor from the category of C^{∞} manifolds and C^{∞} maps to the category of bundles and bundle maps which is naturally equivalent to $(\varepsilon^n, \text{old } f_)$ on Euclidean spaces, and to the restriction of the functor on open submanifolds.

Let (x, U) be a coordinate system on M . Then we have the following string of equivalences. Two of them, which are denoted by the same symbol \simeq , are the equivalences mentioned in condition (3). Let α_x denote the composition $\alpha_x = (i^n|_{x(U)}) \circ \simeq \circ x_* \circ (\simeq)^{-1}$.

$$(TM)|U \xleftarrow{\simeq} TU \xrightarrow{x_*} T(x(U)) \xrightarrow{\simeq} (T\mathbb{R}^n)|_{x(U)} \xrightarrow{i^n|_{x(U)}} \varepsilon^n(\mathbb{R}^n)|_{x(U)}$$

α_x

Similarly, using equivalence \simeq' for T' , we can define β_x .

$$(T'M)|U \xleftarrow{\simeq'} T'U \xrightarrow{x'_*} T'(x(U)) \xrightarrow{\simeq'} (T'\mathbb{R}^n)|_{x(U)} \xrightarrow{i'^n|_{x(U)}} \varepsilon^n(\mathbb{R}^n)|_{x(U)}$$

β_x

Then

$$\beta_x^{-1} \circ \alpha_x: (TM)|U \rightarrow (T'M)|U$$

is an equivalence, so it takes the fibre of TM over p isomorphically to the fibre of $T'M$ over p for each $p \in U$. Our main task is to show that this isomorphism between the fibres over p is independent of the coordinate system (x, U) . This will be done in three stages.

(I) Suppose $V \subset U$ is open and $y = x|_V$. We will need to name all the inclusion maps

$$i: U \rightarrow M$$

$$\bar{i}: V \rightarrow M$$

$$j: V \rightarrow U$$

$$k: y(V) \rightarrow x(U).$$

To compare α_x and α_y , consider the following diagram.

$$(1) \quad \begin{array}{ccccccc} (TM)|U & \xleftarrow{\simeq} & TU & \xrightarrow{x_*} & T(x(U)) & \xrightarrow{\simeq} & (T\mathbb{R}^n)|_{x(U)} \xrightarrow{i^n|_{x(U)}} \varepsilon^n(\mathbb{R}^n)|_{x(U)} \\ \uparrow \subset & \textcircled{1} & \uparrow j_* & \textcircled{2} & \uparrow k_* & \textcircled{3} & \uparrow \subset \textcircled{4} \uparrow \subset \\ (TM)|V & \xleftarrow{\simeq} & TV & \xrightarrow{y_*} & T(y(V)) & \xrightarrow{\simeq} & (T\mathbb{R}^n)|_{y(V)} \xrightarrow{i^n|_{y(V)}} \varepsilon^n(\mathbb{R}^n)|_{y(V)} \end{array}$$

Each of the four squares in this diagram commutes. To see this for square ①, we enlarge it, as shown below. The two triangles on the left commute by condition (3) for TM , and the one on the right commutes because $i \circ j = \bar{i}$.

$$\begin{array}{ccc}
 (TM)|U & & \\
 \downarrow \subset & \swarrow \cong & \\
 & TU & \\
 & \swarrow i_* & \uparrow j_* \\
 TM & & TV \\
 \uparrow \subset & \swarrow \bar{i}_* & \\
 (TM)|V & &
 \end{array}$$

Square ② commutes because $k \circ y = x \circ j$. Square ③ commutes for the same reason as square ①; the inclusions $x(U) \rightarrow \mathbb{R}^n$ and $y(V) \rightarrow \mathbb{R}^n$ come into play. Square ④ obviously commutes. Chasing through diagram (1) now shows that the following commutes.

$$\begin{array}{ccc}
 (TM)|U & \xrightarrow{\alpha_x} & \varepsilon^n(\mathbb{R}^n)|x(U) \\
 \uparrow \subset & & \uparrow \subset \\
 (TM)|V & \xrightarrow{\alpha_y} & \varepsilon^n(\mathbb{R}^n)|y(V)
 \end{array}$$

This means that for $p \in V$, the isomorphism α_y between the fibres over p is the same as α_x . Clearly the same is true for β_x and β_y , since our proof used only properties (1), (2), and (3), not the explicit construction of TM . Thus $\beta_y^{-1} \circ \alpha_y = \beta_x^{-1} \circ \alpha_x$ on the fibres over p , for every $p \in V$.

(II) We now need a Lemma which applies to both TM and $T'M$. Again, it will be proved for TM (where it is actually obvious), using only properties (1), (2), and (3), so that it is also true for $T'M$.

LEMMA. If $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ are open, and $f: A \rightarrow B$ is C^∞ , then the following diagram commutes.

$$\begin{array}{ccccc} TA & \xrightarrow{\simeq} & (T\mathbb{R}^n)|_A & \xrightarrow{i^n|_A} & \varepsilon^n(\mathbb{R}^n)|_A \\ f_* \downarrow & & & & \downarrow \text{old } f_* \\ TB & \xrightarrow{\simeq} & (T\mathbb{R}^m)|_B & \xrightarrow{i^m|_B} & \varepsilon^m(\mathbb{R}^m)|_B \end{array}$$

PROOF. Case 1. There is a map $\bar{f}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $\bar{f} = f$ on A . Consider the following diagram, where $i: A \rightarrow \mathbb{R}^n$ and $j: B \rightarrow \mathbb{R}^m$ are the inclusion maps.

$$\begin{array}{ccccccc} & & & (T\mathbb{R}^n)|_A & \xrightarrow{i^n|_A} & \varepsilon^n(\mathbb{R}^n)|_A & \\ & \nearrow \simeq & & \downarrow \subset & & \downarrow \subset & \\ TA & \xrightarrow{i_*} & T\mathbb{R}^n & \xrightarrow{i^n} & \varepsilon^n(\mathbb{R}^n) & & \\ f_* \downarrow & & \downarrow \bar{f}_* & & \downarrow \text{old } \bar{f}_* & & \\ TB & \xrightarrow{j_*} & T\mathbb{R}^m & \xrightarrow{i^m} & \varepsilon^m(\mathbb{R}^m) & & \\ & \searrow \simeq & & \uparrow \subset & & \uparrow \subset & \\ & & (T\mathbb{R}^m)|_B & \xrightarrow{i^m|_B} & \varepsilon^m(\mathbb{R}^m)|_B & & \end{array}$$

Everything in this diagram obviously commutes. This implies that the two compositions

$$TA \xrightarrow{\simeq} (T\mathbb{R}^n)|_A \xrightarrow{i^n|_A} \varepsilon^n(\mathbb{R}^n)|_A \xrightarrow{\subset} \varepsilon^n(\mathbb{R}^n) \xrightarrow{\text{old } \bar{f}_*} \varepsilon^m(\mathbb{R}^m)$$

and

$$TA \xrightarrow{f_*} TB \xrightarrow{\simeq} (T\mathbb{R}^m)|_B \xrightarrow{i^m|_B} \varepsilon^m(\mathbb{R}^m)|_B \xrightarrow{\subset} \varepsilon^m(\mathbb{R}^m)$$

are equal and this proves the Lemma in Case 1, since the maps “old \bar{f}_* ” and “old f_* ” are equal on A .

Case 2. General case. For each $p \in A$, we want to show that two maps are the same on the fibre over p . Now there is a map $\bar{f}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $\bar{f} = f$ on an open set A' , where $p \in A' \subset A$. We then have the following diagram, where every \simeq comes from the fact that some set is an open submanifold of another,

and $i: A' \rightarrow A$ is the inclusion map.

$$\begin{array}{ccccc}
 & TA & \xrightarrow{\quad} & (T\mathbb{R}^n)|_A & \xrightarrow{i^n|_A} & \varepsilon^n(\mathbb{R}^n)|_A \\
 & \uparrow i_* & \nwarrow \subset & \uparrow \subset & & \uparrow \subset \\
 & (TA)|_{A'} & \xrightarrow{\quad} & (T\mathbb{R}^n)|_{A'} & \xrightarrow{i^n|_{A'}} & \varepsilon^n(\mathbb{R}^n)|_{A'} \\
 \textcircled{1} & \uparrow f_* & \nwarrow \cong & \uparrow \subset & & \uparrow \subset \\
 TA' & \xrightarrow{\quad} & (T\mathbb{R}^n)|_{A'} & \xrightarrow{i^n|_{A'}} & \varepsilon^n(\mathbb{R}^n)|_{A'} & \textcircled{5} \\
 \downarrow (f|_{A'})_* & & \downarrow \text{old } (f|_{A'})_* & & \downarrow \text{old } \bar{f}_* & \\
 TB & \xrightarrow{\quad} & (T\mathbb{R}^m)|_B & \xrightarrow{i^m|_B} & \varepsilon^m(\mathbb{R}^m)|_B
 \end{array}$$

$\textcircled{2}$ $\textcircled{3}$ $\textcircled{4}$

Boxes $\textcircled{1}$, $\textcircled{3}$, and $\textcircled{5}$ obviously commute, and $\textcircled{4}$ commutes by *Case 1*. To see that square $\textcircled{2}$ (which has a triangle within it) commutes, we imbed it in a larger diagram, in which $j: A \rightarrow \mathbb{R}^n$ is the inclusion map, and other maps have also been named, for ease of reference.

$$\begin{array}{ccc}
 & T\mathbb{R}^n & \\
 j_* \nearrow & & \nwarrow \subset (\nu) \\
 TA & \xrightarrow{\cong (\lambda)} & (T\mathbb{R}^n)|_A \\
 i_* \uparrow & & \uparrow \subset (\mu) \\
 TA' & \xrightarrow{\cong (\kappa)} & (T\mathbb{R}^n)|_{A'}
 \end{array}$$

To prove that $\lambda \circ i_* = \mu \circ \kappa$, it suffices to prove that

$$\nu \circ \lambda \circ i_* = \nu \circ \mu \circ \kappa,$$

since ν is one-one. Thus it suffices to prove $j_* \circ i_* = \nu \circ \mu \circ \kappa$, which amounts to proving commutativity of the following diagram.

$$\begin{array}{ccc}
 & T\mathbb{R}^n & \\
 (j \circ i)_* \nearrow & & \nwarrow \subset \\
 TA' & \xrightarrow{\cong} & (T\mathbb{R}^n)|_{A'}
 \end{array}$$

Since $j \circ i$ is just the inclusion of A' in \mathbb{R}^n , this does commute.

Commutativity of diagram (2) shows that the composition

$$TA \xrightarrow{f_*} TB \xrightarrow{\cong} (T\mathbb{R}^m)|_B \xrightarrow{t^m|_B} \varepsilon^m(\mathbb{R}^m)|_B$$

coincides, on the subset $(TA)|_{A'}$, with the composition

$$TA \xrightarrow{\cong} (T\mathbb{R}^n)|_A \xrightarrow{t^n|_A} \varepsilon^n(\mathbb{R}^n)|_A \xrightarrow{\text{old } \tilde{f}_*} \varepsilon^m(\mathbb{R}^m)|_B,$$

and on A' we can replace “old \tilde{f}_* ” by “old f_* ”. In other words, the two compositions are equal in a neighborhood of any $p \in A$, and are thus equal, which proves the Lemma.

(III) Now suppose (x, U) and (y, V) are any two coordinate systems with $p \in U \cap V$. To prove that $\beta_y^{-1} \circ \alpha_y$ and $\beta_x^{-1} \circ \alpha_x$ induce the same isomorphism on the fibre of TM at p , we can assume without loss of generality that $U = V$, because part (I) applies to x and $x|_{U \cap V}$, as well as to y and $y|_{U \cap V}$.

Assuming $U = V$, we have the following diagram.

$$(3) \quad \begin{array}{ccccc} & & T(x(U)) & \xrightarrow{\cong} & (T\mathbb{R}^n)|_{x(U)} & \xrightarrow{t^n|_{x(U)}} & \varepsilon^n(\mathbb{R}^n)|_{x(U)} \\ & \nearrow x_* & \downarrow (y \circ x^{-1})_* & & \downarrow \text{old } (y \circ x^{-1})_* & & \downarrow \\ (TM)|_U & \xleftarrow{\cong} & TU & & & & \\ & \searrow y_* & \downarrow & & & & \\ & & T(y(U)) & \xrightarrow{\cong} & (T\mathbb{R}^n)|_{y(U)} & \xrightarrow{t^n|_{y(U)}} & \varepsilon^n(\mathbb{R}^n)|_{y(U)} \end{array}$$

The triangle obviously commutes, and the rectangle commutes by part (II). Diagram (3) thus shows that

$$\alpha_y = \text{old } (y \circ x^{-1})_* \circ \alpha_x.$$

Exactly the same result holds for T' :

$$\beta_y = \text{old } (y \circ x^{-1})_* \circ \beta_x.$$

The desired result $\beta_y^{-1} \circ \alpha_y = \beta_x^{-1} \circ \alpha_x$ follows immediately.

Now that we have a well-defined bundle map $TM \rightarrow T'M$ (the union of all $\beta_x^{-1} \circ \alpha_x$), it is clearly an equivalence e_M . The proof that $e_N \circ f_* = f_{\sharp} \circ e_M$ is left as a masochistic exercise for the reader. ♦

PROBLEMS

1. Let M be any set, and $\{(x_i, U_i)\}$ a sequence of one-one functions $x_i: U_i \rightarrow \mathbb{R}^n$ with $U_i \subset M$ and $x(U_i)$ open in \mathbb{R}^n , such that each

$$x_j \circ x_i^{-1}: x_i(U_i \cap U_j) \rightarrow x_j(U_i \cap U_j)$$

is continuous. It would seem that M ought to have a metric which makes each U_i open and each x_i a homeomorphism. Actually, this is not quite true:

(a) Let $M = \mathbb{R} \cup \{*\}$, where $* \notin \mathbb{R}$. Let $U_1 = \mathbb{R}$ and $x_1: U_1 \rightarrow \mathbb{R}$ be the identity, and let $U_2 = \mathbb{R} - \{0\} \cup \{*\}$, with $x_2: U_2 \rightarrow \mathbb{R}$ defined by

$$\begin{aligned} x_2(a) &= a, & a \neq 0, * \\ x_2(*) &= 0. \end{aligned}$$

Show that there is no metric on M of the required sort, by showing that every neighborhood of 0 would have to intersect every neighborhood of *. Nevertheless, we can find on M a pseudometric ρ (a function $\rho: M \times M \rightarrow \mathbb{R}$ with all properties for a metric except that $\rho(p, q)$ may be 0 for $p \neq q$) such that ρ is a metric on each U_i and each x_i is a homeomorphism:

(b) If $A \subset \mathbb{R}^n$ is open, then there is a sequence A_1, A_2, A_3, \dots of open subsets of A such that every open subset of A is a union of certain A_i 's.

(c) There is a sequence of continuous functions $f_i: A \rightarrow [0, 1]$, with support $f_i \subset A$, which "separates points and closed sets": if C is closed and $p \in A - C$, then there is some f_i with $f_i(p) \notin f_i(A \cap C)$. *Hint:* First arrange in a sequence all pairs (A_i, A_j) of part (b) with $\overline{A_i} \subset A_j$.

(d) Let $f_{i,j}$, $j = 1, 2, 3, \dots$ be such a sequence for each open set $x_i(U_i)$. Define $g_{i,j}: M \rightarrow [0, 1]$ by

$$g_{i,j}(p) = \begin{cases} f_{i,j}(p) & p \in U_i \\ 0 & p \notin U_i. \end{cases}$$

Arrange all $g_{i,j}$ in a single sequence G_1, G_2, G_3, \dots , let d be a bounded metric on \mathbb{R} , and define ρ on M by

$$\rho(p, q) = \sum_{i=1}^{\infty} \frac{1}{2^i} d(G_i(p), G_i(q)).$$

Show that ρ is the required pseudometric.

(e) Suppose that for every $p, q \in M$ there is a U_i and U_j with $p \in U_i$ and $q \in U_j$ and open sets $B_i \subset x_i(U_i)$ and $B_j \subset x_j(U_j)$ so that $p \in x_i^{-1}(B_i)$, $q \in x_j^{-1}(B_j)$, and $x_i^{-1}(B_i) \cap x_j^{-1}(B_j) = \emptyset$. Show that ρ is actually a metric on M .

2. (a) Suppose (x, U) and (y, V) are two coordinate systems, giving rise to two maps on TM ,

$$\begin{aligned} t_x: \pi^{-1}(U) &\rightarrow U \times \mathbb{R}^n, & [x, v]_q &\mapsto (q, v), \\ t_y: \pi^{-1}(V) &\rightarrow V \times \mathbb{R}^n, & [y, w]_q &\mapsto (q, w). \end{aligned}$$

Show that in $\pi^{-1}(U \cap V)$ the sets of the form $t_x^{-1}(A)$ for $A \subset U \times \mathbb{R}^n$ open are exactly the sets of the form $t_y^{-1}(B)$ for $B \subset V \times \mathbb{R}^n$ open.

(b) Show that if there is a metric on TM such that t_{x_i} is a homeomorphism for a collection (x_i, U_i) with $M = \bigcup_i U_i$, then all t_x are homeomorphisms.

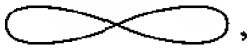
(c) Conclude from Problem 1 that there is a metric on TM which makes each t_x a homeomorphism.

3. Show that in the definition of an equivalence it suffices to assume that the map $E_1 \rightarrow E_2$ is continuous. (To prove the inverse continuous, note that locally it is just a map $U \times \mathbb{R}^n \rightarrow U \times \mathbb{R}^n$).

4. Show that in the definition of a bundle map, continuity of $f: B_1 \rightarrow B_2$ follows automatically from continuity of $\tilde{f}: E_1 \rightarrow E_2$.

5. A weak equivalence between two bundles over the same base space B is a bundle map (\tilde{f}, f) where \tilde{f} is an isomorphism on each fibre, and f is a homeomorphism of B onto itself. Find two inequivalent, but weakly equivalent, bundles over the following base spaces:

(i) the disjoint union of two circles,

(ii) a figure eight ,

(iii) the torus.

6. Given a bundle map (\tilde{f}, f) , show that $\tilde{f} = g \circ h$ where g and h are continuous maps such that h takes fibres linearly to fibres, while g is an isomorphism on each fibre.

7. (a) Show that for any bundle $\pi: E \rightarrow B$, the map $s: B \rightarrow E$ with $s(p)$ the 0 vector of $\pi^{-1}(p)$ is a section.

(b) Show that an n -plane bundle ξ is trivial if and only if there are n sections s_1, \dots, s_n which are everywhere linearly independent, i.e., $s_1(p), \dots, s_n(p) \in \pi^{-1}(p)$ are linearly independent for all $p \in B$.

(c) Show that locally every n -plane bundle has n linearly independent sections.

8. (a) Check that \sim_p is an equivalence relation on the set of pairs (x, v) .

(b) Check that the definition of f_* is independent of the coordinate systems x and y which are used.

(c) Check the remaining details in Theorem 1.

9. (a) Show that the correspondence between TM and equivalence classes of curves under which $[x, v]_p$ corresponds to the \approx equivalence class of $x^{-1} \circ \gamma$, for γ a curve in \mathbb{R}^n with $\gamma'(0) = v$, makes f_* correspond to $f_\#$.

(b) Show that under the correspondence $[x, a]_p \mapsto \sum_i a^i \partial/\partial x^i|_p$, the map f_* can be defined by

$$[f_*(\ell)](g) = \ell(g \circ f).$$

10. If V is a finite dimensional vector space over \mathbb{R} , define a C^∞ structure on V and a homeomorphism from $V \times V$ to TV which is independent of choice of bases. As in the case of \mathbb{R}^n , for $v, w \in V$ we will denote by $v_w \in V_w$ the vector corresponding to (w, v) .

11. If $g: \mathbb{R} \rightarrow \mathbb{R}$ is C^∞ show that

$$g(x) = g(0) + g'(0)x + x^2h(x)$$

for some C^∞ function $h: \mathbb{R} \rightarrow \mathbb{R}$.

12. (a) Let \mathcal{F}_p be the set of all C^∞ functions $f: M \rightarrow \mathbb{R}$ with $f(p) = 0$, and let $\ell: \mathcal{F}_p \rightarrow \mathbb{R}$ be a linear operator with $\ell(fg) = 0$ for all $f, g \in \mathcal{F}_p$. Show that ℓ has a unique extension to a derivation.

(b) Let W be the vector subspace of \mathcal{F}_p generated by all products fg for $f, g \in \mathcal{F}_p$. Show that the vector space of all derivations at p is isomorphic to the dual space $(\mathcal{F}_p/W)^*$.

(c) Since $(\mathcal{F}_p/W)^*$ has dimension $n = \text{dimension of } M$, the same must be true of \mathcal{F}_p/W . If x is a coordinate system with $x(p) = 0$, show that $x^1 + W, \dots, x^n + W$ is a basis for \mathcal{F}_p/W (use Lemma 2). The situation is quite different for C^1 functions, as the next problem shows.

13. (a) Let V be the vector space of all C^1 functions $f: \mathbb{R} \rightarrow \mathbb{R}$ with $f(0) = 0$, and let W be the subspace generated by all products. Show that $\lim_{x \rightarrow 0} f(x)/x^2$ exists for all $f \in W$.

(b) For $0 < \varepsilon < 1$, let

$$f_\varepsilon(x) = \begin{cases} x^{1+\varepsilon} & x \geq 0 \\ 0 & x \leq 0. \end{cases}$$

Show that all f_ε are in V , and that they represent linearly independent elements of V/W .

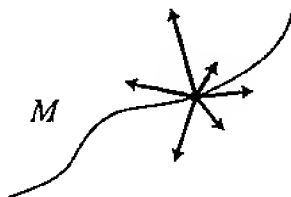
(c) Conclude that $(V/W)^*$ has dimension $c^c = 2^c$.

14. If $f: M \rightarrow N$ and f_* is the 0 map on each fibre, then f is constant on each component of M .

15. (a) A map $f: M \rightarrow N$ is an immersion if and only if f_* is one-one on each fibre of TM . More generally, the rank of f at $p \in M$ is the rank of the linear transformation $f_*: M_p \rightarrow N_{f(p)}$.

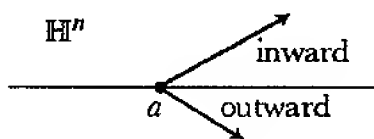
(b) If $f \circ g = f$, where g is a diffeomorphism, then the rank of $f \circ g$ at a equals the rank of f at $g(a)$. (Compare with Problem 2-33(d).)

16. (a) If M is a manifold-with-boundary, the tangent bundle TM is defined exactly as for M ; elements of M_p are \sim_p equivalence classes of pairs (x, v) . Although x takes a neighborhood of $p \in \partial M$ onto \mathbb{H}^n , rather than \mathbb{R}^n , the vectors v still run through \mathbb{R}^n , so M_p still has tangent vectors “pointing in all directions”. If $p \in \partial M$ and $x: U \rightarrow \mathbb{H}^n$ is a coordinate system around p , then



$x_*^{-1}(\mathbb{R}^{n-1}_{x(p)}) \subset M_p$ is a subspace. Show that this subspace does not depend on the choice of x ; in fact, it is $i_*(\partial M)_p$, where $i: \partial M \rightarrow M$ is the inclusion.

(b) Let $a \in \mathbb{R}^{n-1} \times \{0\} \subset \mathbb{H}^n$. A tangent vector in \mathbb{H}^n_a is said to point “inward” if, under the identification of $T\mathbb{H}^n$ with $\varepsilon^n(\mathbb{H}^n)$, the vector is (a, v) where $v^n > 0$. A vector $v \in M_p$ which is not in $i_*(\partial M)_p$ is said to point “inward” if



$x_*(v) \in \mathbb{H}^n_{x(p)}$ points inward. Show that this definition does not depend on the coordinate system x .

(c) Show that if M has an orientation μ , then ∂M has a unique orientation $\partial\mu$ such that $[v_1, \dots, v_{n-1}] = (\partial\mu)_p$ if and only if $[w, i_*v_1, \dots, i_*v_{n-1}] = \mu_p$ for every outward pointing $w \in M_p$.

(d) If μ is the usual orientation of \mathbb{H}^n , show that $\partial\mu$ is $(-1)^n$ times the usual orientation of $\mathbb{R}^{n-1} = \partial\mathbb{H}^n$. (The reason for this choice will become clear in Chapter 8.)

(e) Suppose we are in the setup of Problem 2-14. Define $g: \partial M \times [0, 1) \rightarrow \partial N \times [0, 1)$ by $g(p, t) = (f(p), t)$. Show that TP is obtained from $TM \cup TN$

by identifying

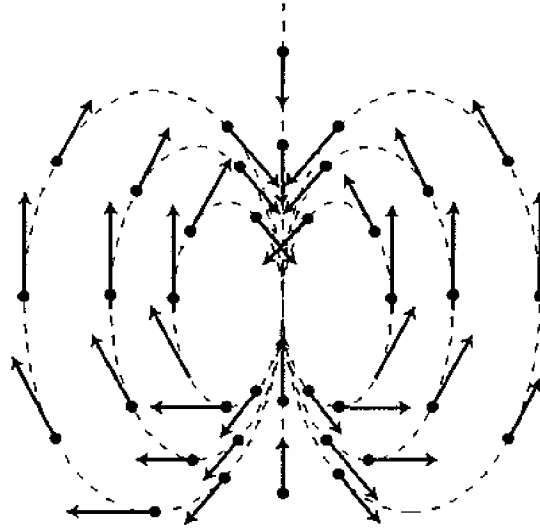
$$v \in (\partial M)_p \quad \text{with} \quad (\beta^{-1})_* g_* \alpha_*(v) \in (\partial N)_{f(p)}.$$

(f) If M and N have orientations μ and ν and $f: (\partial M, \partial\mu) \rightarrow (\partial N, \partial\nu)$ is orientation-reversing, show that P has an orientation which agrees with μ and ν on $M \subset P$ and $N \subset P$.

(g) Suppose M is S^2 with two holes cut out, and N is $[0, 1] \times S^1$. Let f be a diffeomorphism from M to N which is orientation preserving on one copy of S^1 and orientation reversing on the other. What is the resulting manifold P ?

17. Show that $T\mathbb{P}^2$ is homeomorphic to the space obtained from $T(S^2, i)$ by identifying $(p, v) \in (S^2, i)_p$ with $(-p, -v) \in (S^2, i)_{-p}$.

18. Although there is no everywhere non-zero vector field on S^2 , there is one on $S^2 - \{(0, 0, 1)\}$, which is diffeomorphic to \mathbb{R}^2 . Show that such a vector field can be picked so that near $(0, 0, 1)$ the vector field looks like the following picture (a “magnetic dipole”):



19. Suppose we have a “multiplication” map $(a, b) \mapsto a \cdot b$ from $\mathbb{R}^n \times \mathbb{R}^n$ to \mathbb{R}^n that makes \mathbb{R}^n into a (non-associative) division algebra. That is,

$$(a_1 + a_2) \cdot b = a_1 \cdot b + a_2 \cdot b$$

$$a \cdot (b_1 + b_2) = a \cdot b_1 + a \cdot b_2$$

$$\lambda(a \cdot b) = (\lambda a) \cdot b = a \cdot (\lambda b) \quad \text{for } \lambda \in \mathbb{R}$$

$$a \cdot (1, 0, \dots, 0) = a$$

and there are no zero divisors:

$$a, b \neq 0 \implies ab \neq 0.$$

(For example, for $n = 1$, we can use ordinary multiplication, and for $n = 2$ we can use “complex multiplication”, $(a, b) \cdot (c, d) = (ac - bd, ad + bc)$.) Let e_1, \dots, e_n be the standard basis of \mathbb{R}^n .

- (a) Every point in S^{n-1} is $a \cdot e_1$ for a unique $a \in \mathbb{R}^n$.
- (b) If $a \neq 0$, then $a \cdot e_1, \dots, a \cdot e_n$ are linearly independent.
- (c) If $p = a \cdot e_1 \in S^{n-1}$, then the projection of $a \cdot e_2, \dots, a \cdot e_n$ on $(S^{n-1}, i)_p$ are linearly independent.
- (d) Multiplication by a is continuous.
- (e) TS^{n-1} is trivial.
- (f) $T\mathbb{P}^{n-1}$ is trivial.

The tangent bundles TS^3 and TS^7 are both trivial. Multiplications with the required properties on \mathbb{R}^4 and \mathbb{R}^8 are provided by the “quaternions” and “Cayley numbers”, respectively; the quaternions are not commutative and the Cayley numbers are not even associative. It is a classical theorem that the reals, complexes, and quaternions are the only associative examples. For a simple proof, see R. S. Palais, *The Classification of Real Division Algebras*, Amer. Math. Monthly 75 (1968), 366–368. J. F. Adams has proved, using methods of algebraic topology, that $n = 1, 2, 4$, or 8 .

[Incidentally, non-existence of zero divisors immediately implies that for $a \neq 0$ there is some b with $ab = (1, 0, \dots, 0)$ and b' with $b'a = (1, 0, \dots, 0)$. If the multiplication is associative it follows easily that $b = b'$, so that we always have multiplicative inverses. Conversely, this condition implies that there are no zero divisors if the multiplication is associative; otherwise it suffices to assume the existence of a *unique* b with $a \cdot b = b \cdot a = (1, 0, \dots, 0)$.]

20. (a) Consider the space obtained from $[0, 1] \times \mathbb{R}^n$ by identifying $(0, v)$ with $(1, Tv)$, where $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a vector space isomorphism. Show that this can be made into the total space of a vector bundle over S^1 (a generalized Möbius strip).

(b) Show that the resulting bundle is orientable if and only if T is orientation preserving.

21. Show that for $p \in S^2$, the vector $p_p \in \mathbb{R}^3_p$ is not in $i_*(S^2_p)$ by showing that the inner product $\langle p, c'(t) \rangle = 0$ for all curves c with $c(0) = p$ and $|c(t)| = 1$ for all t . (Recall that

$$\langle f, g \rangle'(t) = (f'(t))^t, g(t) \rangle + \langle f(t), g'(t)^t \rangle,$$

where t denotes the transpose; see *Calculus on Manifolds*, pg. 23.)

22. Let M be a C^∞ manifold. Suppose that $(TM)|_A$ is trivial whenever $A \subset M$ is homeomorphic to S^1 . Show that M is orientable. *Hint:* An arc c from

$p_0 \in M$ to $p \in M$ is contained in some such A so $(TM)|_c$ is trivial. Thus one can “transport” the orientation of M_{p_0} to M_p . It must be checked that this is independent of the choice of c . First consider pairs c, c' which meet only at p_0 and p . The general, possibly quite messy, case can be treated by breaking up c into small pieces contained in coordinate neighborhoods.

Remark: Using results from the Addendum to Chapter 9, together with Problem 29, we can conclude that a neighborhood of some $S^1 \subset M$ is non-orientable if M is non-orientable.

The next two problems deal with important constructions associated with vector bundles.

23. (a) Suppose $\xi = \pi: E \rightarrow X$ is a bundle and $f: Y \rightarrow X$ is a continuous map. Let $E' \subset Y \times E$ be the set of all (y, e) with $f(y) = \pi(e)$, define $\pi': E' \rightarrow Y$ by $\pi'(y, e) = y$, and define $\tilde{f}: E' \rightarrow E$ by $\tilde{f}(y, e) = e$. A vector space structure can be defined on

$$\pi'^{-1}(y) = \{(y, e): e \in \pi^{-1}(f(y))\}$$

by using the vector space structure on $\pi^{-1}(f(y))$. Show that $\pi': E' \rightarrow Y$ is a bundle, and (\tilde{f}, f) a bundle map which is an isomorphism on each fibre. This bundle is denoted by $f^*(\xi)$, and is called the bundle induced (from ξ) by f .

(b) Suppose we have another bundle $\xi'' = \pi'': E'' \rightarrow Y$ and a bundle map (\tilde{f}, f) from ξ'' to ξ which is an isomorphism on each fibre. Show that $\xi'' \simeq \xi' = f^*(\xi)$. *Hint:* Map $e \in E''$ to $(\pi''(e), \tilde{f}(e)) \in E'$.

(c) If $g: Z \rightarrow Y$, then $(f \circ g)^*(\xi) \simeq g^*(f^*(\xi))$.

(d) If $A \subset X$ and $i: A \rightarrow X$ is the inclusion map, then $i^*(\xi) \simeq \xi|_A$.

(e) If ξ is orientable, then $f^*(\xi)$ is also orientable.

(f) Give an example where ξ is non-orientable, but $f^*(\xi)$ is orientable.

(g) Let $\xi = \pi: E \rightarrow B$ be a vector bundle. Since $\pi: E \rightarrow B$ is a continuous map from a space to the base space B of ξ , the symbol $\pi^*(\xi)$ makes sense. Show that if ξ is not orientable, then $\pi^*(\xi)$ is not orientable.

24. (a) Given an n -plane bundle $\xi = \pi: E \rightarrow B$ and an m -plane bundle $\eta = \pi': E' \rightarrow B$, let $E'' \subset E \times E'$ be the set of all pairs (e, e') with $\pi(e) = \pi'(e')$. Let $\pi''(e, e') = \pi(e) = \pi'(e')$. Show that $\pi'': E'' \rightarrow B$ is an $(n + m)$ -plane bundle. It is called the Whitney sum $\xi \oplus \eta$ of ξ and η ; the fibre of $\xi \oplus \eta$ over p is the direct sum $\pi^{-1}(p) \oplus \pi'^{-1}(p)$.

(b) If $f: Y \rightarrow B$, show that $f^*(\xi \oplus \eta) \simeq f^*(\xi) \oplus f^*(\eta)$.

- (c) Given bundles $\xi_i = \pi_i: E_i \rightarrow B_i$, define $\pi: E_1 \times E_2 \rightarrow B_1 \times B_2$ by $\pi(e_1, e_2) = (\pi_1(e_1), \pi_2(e_2))$. Show that this is a bundle $\xi_1 \times \xi_2$ over $B_1 \times B_2$.
- (d) If $\Delta: B \rightarrow B \times B$ is the “diagonal map”, $\Delta(x) = (x, x)$, show that $\xi \oplus \eta \simeq \Delta^*(\xi \times \eta)$.
- (e) If ξ and η are orientable, show that $\xi \oplus \eta$ is orientable.
- (f) If ξ is orientable, and η is non-orientable, show that $\xi \oplus \eta$ is also non-orientable.
- (g) Define a “natural” orientation on $V \oplus V$ for any vector space V , and use this to show that $\xi \oplus \xi$ is always orientable.
- (h) If X is a “figure eight” (c.f. Problem 5), find two non-orientable 1-plane bundles ξ and η over X such that $\xi \oplus \eta$ is also non-orientable.

25. (a) If $\pi: E \rightarrow M$ is a C^∞ vector bundle, then π_* has maximal rank at each point, and each fibre $\pi^{-1}(p)$ is a C^∞ submanifold of E .

(b) The 0-section of E is a submanifold, carried diffeomorphically onto B by π .

26. (a) If M and N are C^∞ manifolds, and π_M [or π_N] : $M \times N \rightarrow M$ [or N] is the projection on M [or N], then $T(M \times N) \simeq \pi_M^*(TM) \oplus \pi_N^*(TN)$.

(b) If M and N are orientable, then $M \times N$ is orientable.

(c) If $M \times N$ is orientable, then both M and N are orientable.

27. Show that the Jacobian matrix of $y_* \circ (x_*)^{-1}$ is of the form

$$\begin{pmatrix} D_j y^i \circ x^{-1} & 0 \\ \times & D_j y^i \circ x^{-1} \end{pmatrix}.$$

This shows that the manifold TM is *always orientable*, i.e., the bundle $T(TM)$ is orientable. (Here is a more conceptual formulation: for $v \in TM$, the orientation for $(TM)_v$ can be defined as

$$\left[\frac{\partial}{\partial(x^1 \circ \pi)} \Big|_v, \dots, \frac{\partial}{\partial(x^n \circ \pi)} \Big|_v, \frac{\partial}{\partial \dot{x}^1} \Big|_v, \dots, \frac{\partial}{\partial \dot{x}^n} \Big|_v \right];$$

the form of $y_* \circ (x_*)^{-1}$ shows that this orientation is independent of the choice of x .) A different proof that TM is orientable is given in Problem 29.

28. (a) Let (x, U) be a coordinate system on M with $x(p) = 0$ and let $v \in M_p$ be $\sum_{i=1}^n a^i \partial/\partial x^i|_p$. Consider the curve c in TM defined by

$$c(t) = v + t \frac{\partial}{\partial x^i} \Big|_p.$$

Show that

$$\frac{dc}{dt}(0) = \frac{\partial}{\partial \dot{x}^i} \Big|_v.$$

(b) Find a curve whose tangent vector at 0 is $\partial/\partial(x^i \circ \pi)|_v$.

29. This problem requires some familiarity with the notion of exact sequences

(c.f. Chapter 11). A sequence of bundle maps $E_1 \xrightarrow{\tilde{f}} E_2 \xrightarrow{\tilde{g}} E_3$ with $f = g =$ identity of B is exact if at each fibre it is exact as a sequence of vector space maps.

(a) If $\xi = \pi: E \rightarrow B$ is a C^∞ vector bundle, show that there is an exact sequence

$$0 \rightarrow \pi^*(\xi) \rightarrow TE \rightarrow \pi^*(TB) \rightarrow 0.$$

Hint: (1) An element of the total space of $\pi^*(\xi)$ is a pair of points in the same fibre, which determines a tangent vector of the fibre. (2) Map $X \in (TE)_e$ to $(e, \pi_* X)$.

(b) If $0 \rightarrow E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow 0$ is exact, then each bundle E_i is orientable if the other two are.

(c) $T(TM)$ is always orientable.

(d) If $\pi: E \rightarrow M$ is not orientable, then the manifold E is not orientable. (This is why the proof that the Möbius strip is a non-orientable manifold is so similar to the proof that the Möbius bundle over S^1 is not orientable.)

The next two Problems contain more information about the groups introduced in Problem 2-33. In addition to being used in Problem 32, this information will all be important in Chapter 10.

30. (a) Let $p_0 \in S^{n-1}$ be the point $(0, \dots, 0, 1)$. For $n \geq 2$ define $f: \text{SO}(n) \rightarrow S^{n-1}$ by $f(A) = A(p_0)$. Show that f is continuous and open. Show that $f^{-1}(p_0)$ is homeomorphic to $\text{SO}(n-1)$, and then show that $f^{-1}(p)$ is homeomorphic to $\text{SO}(n-1)$ for all $p \in S^{n-1}$.

(b) $\text{SO}(1)$ is a point, so it is connected. Using part (a), and induction on n , prove that $\text{SO}(n)$ is connected for all $n \geq 1$.

(c) Show that $\text{O}(n)$ has exactly two components.

31. (a) If $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear transformation, $T^*: \mathbb{R}^n \rightarrow \mathbb{R}^n$, the adjoint of T , is defined by $(T^*v, w) = (v, Tw)$ (for each v , the map $w \mapsto (v, Tw)$ is linear, so it is $w \mapsto (T^*v, w)$ for a unique T^*v). If A is the matrix of T with respect to the usual basis, show that the matrix of T^* is the transpose A^t .

(b) A linear transformation $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is self-adjoint if $T = T^*$, so that $(Tv, w) = (v, Tw)$ for all $v, w \in \mathbb{R}^n$. If A is the matrix of T with respect to the standard basis, then T is self-adjoint if and only if A is symmetric, $A^t = A$. It is a standard theorem that a symmetric A can be written as CDC^{-1} for some diagonal matrix D (for an analytic proof, see *Calculus on Manifolds*, pg. 122). Show that C can be chosen orthogonal, by showing that eigenvectors for distinct eigenvalues are orthogonal.

(c) A self-adjoint T (or the corresponding symmetric A) is called **positive semi-definite** if $(Tv, v) \geq 0$ for all $v \in \mathbb{R}^n$, and **positive definite** if $(Tv, v) > 0$ for all $v \neq 0$. Show that a positive definite A is non-singular. *Hint*: Use the Schwarz inequality.

(d) Show that $A^t \cdot A$ is always positive semi-definite.

(e) Show that a positive semi-definite A can be written as $A = B^2$ for some B . (Remember that A is symmetric.)

(f) Show that every $A \in \text{GL}(n, \mathbb{R})$ can be written uniquely as $A = A_1 \cdot A_2$ where $A_1 \in \text{O}(n)$ and A_2 is positive definite. *Hint*: Consider $A^t \cdot A$, and use part (e).

(g) The matrices A_1 and A_2 are continuous functions of A . *Hint*: If $A^{(n)} \rightarrow A$ and $A^{(n)} = A^{(n)}_1 \cdot A^{(n)}_2$, then some subsequence of $\{A^{(n)}_1\}$ converges.

(h) $\text{GL}(n, \mathbb{R})$ is homeomorphic to $\text{O}(n) \times \mathbb{R}^{n(n+1)/2}$ and has exactly two components, $\{A: \det A > 0\}$ and $\{A: \det A < 0\}$. (Notice that this also gives us another way of finding the dimension of $\text{O}(n)$.)

32. Two continuous functions $f_0, f_1: X \rightarrow Y$ are called **homotopic** if there is a continuous function $H: X \times [0, 1] \rightarrow Y$ such that

$$f_i(x) = H(x, i) \quad i = 0, 1.$$

The functions $H_t: X \rightarrow Y$ defined by $H_t(x) = H(x, t)$ may be thought of as a path of functions from $H_0 = f_0$ to $H_1 = f_1$. The map H is called a **homotopy** between f_0 and f_1 .

The notation $f: (X, A) \rightarrow (Y, B)$, for $A \subset X$ and $B \subset Y$, means that $f: X \rightarrow Y$ and $f(A) \subset B$. We call $f_0, f_1: (X, A) \rightarrow (Y, B)$ **homotopic** (as maps from (X, A) to (Y, B)) if there is an H as above such that each $H_t: (X, A) \rightarrow (Y, B)$.

(a) If $A: [0, 1] \rightarrow \text{GL}(n, \mathbb{R})$ is continuous and $H: \mathbb{R}^n \times [0, 1] \rightarrow \mathbb{R}^n$ is defined by $H(x, t) = A(t)(x)$, show that H is continuous, so that H_0 and H_1 are homotopic as maps from $(\mathbb{R}^n, \mathbb{R}^n - \{0\})$ to $(\mathbb{R}^n, \mathbb{R}^n - \{0\})$. Conclude that a non-singular linear transformation $T: (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \rightarrow (\mathbb{R}^n, \mathbb{R}^n - \{0\})$ with $\det T > 0$ is homotopic to the identity map.

(b) Suppose $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is C^∞ and $f(0) = 0$, while $f(\mathbb{R}^n - \{0\}) \subset \mathbb{R}^n - \{0\}$. If $Df(0)$ is non-singular, show that $f: (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \rightarrow (\mathbb{R}^n, \mathbb{R}^n - \{0\})$ is

homotopic to $Df(0): (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \rightarrow (\mathbb{R}^n, \mathbb{R}^n - \{0\})$. *Hint:* Define $H(x, t) = f(tx)$ for $0 < t \leq 1$ and $H(x, 0) = Df(0)(x)$. To prove continuity at points $(x, 0)$, use Lemma 2.

(c) Let U be a neighborhood of $0 \in \mathbb{R}^n$ and $f: U \rightarrow \mathbb{R}^n$ a homeomorphism with $f(0) = 0$. Let $B_r \subset V$ be the open ball with center 0 and radius r , and let $h: \mathbb{R}^n \rightarrow B_r$ be the homeomorphism

$$h(x) = \left(\frac{2r}{\pi} \arctan |x| \right) x;$$

then

$$f \circ h: (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \rightarrow (\mathbb{R}^n, \mathbb{R}^n - \{0\}).$$

We will say that f is **orientation preserving** at 0 if $f \circ h$ is homotopic to the identity map $1: (\mathbb{R}^n, \mathbb{R}^n - \{0\}) \rightarrow (\mathbb{R}^n, \mathbb{R}^n - \{0\})$. Check that this does not depend on the choice of $B_r \subset V$.

(d) For $p \in \mathbb{R}^n$, let $T_p: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be $T_p(q) = p + q$. If $f: U \rightarrow V$ is a homeomorphism, where $U, V \subset \mathbb{R}^n$ are open, we will say that f is **orientation preserving** at p if $T_{-f(p)} \circ f \circ T_p$ is orientation preserving at 0. Show that if M is orientable, then there is a collection \mathcal{C} of charts whose domains cover M such that for every (x, U) and (y, V) in \mathcal{C} , the map $y \circ x^{-1}$ is orientation preserving at $x(p)$ for all $p \in U \cap V$.

(e) Notice that the condition on $y \circ x^{-1}$ in part (d) makes sense even if $y \circ x^{-1}$ is not differentiable. Thus, if M is any (not necessarily differentiable) manifold, we can define M to be **orientable** if there is a collection \mathcal{C} of homeomorphisms $x: U \rightarrow \mathbb{R}^n$ whose domains cover M , such that \mathcal{C} satisfies the condition in part (d). To prove that this definition agrees with the old one we need a fact from algebraic topology: If $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a homeomorphism with $f(0) = 0$ and $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is $T(x^1, \dots, x^n) = (x^1, \dots, x^{n-1}, -x^n)$, then precisely one of f and $T \circ f$ is orientation preserving at 0. Assuming this result, show that if M has such a collection \mathcal{C} of homeomorphisms, then for any C^∞ structure on M the tangent bundle TM is orientable.

33. Let $M^n \subset \mathbb{R}^N$ be a C^∞ n -dimensional submanifold. By a **chord** of M we mean a point of \mathbb{R}^N of the form $p - q$ for $p, q \in M$.

(a) Prove that if $N > 2n + 1$, then there is a vector $v \in S^{N-1}$ such that

- (i) no chord of M is parallel to v ,
- (ii) no tangent plane M_p contains v .

Hint: Consider certain maps from appropriate open subsets of $M \times M$ and TM to S^{N-1} .

(b) Let $\mathbb{R}^{N-1} \subset \mathbb{R}^N$ be the subspace perpendicular to v , and $\pi: \mathbb{R}^N \rightarrow \mathbb{R}^{N-1}$ the corresponding projection. Show that $\pi|_M$ is a one-one immersion. In particular, if M is compact, then $\pi|_M$ is an imbedding.

(c) Every compact C^∞ n -dimensional manifold can be imbedded in \mathbb{R}^{2n+1} .

Note: This is the easy case of Whitney's classical theorem, which gives the same result even for non-compact manifolds (H. Whitney, *Differentiable manifolds*, Ann. of Math. 37 (1935), 645–680). Proofs may be found in Auslander and MacKenzie, *Introduction to Differentiable Manifolds* and Sternberg, *Lectures on Differential Geometry*. In Munkres, *Elementary Differential Topology*, there is a different sort of argument to prove that a not-necessarily-compact n -manifold M can be imbedded in some \mathbb{R}^N (in fact, with $N = (n+1)^2$). Then we may show that M imbeds in \mathbb{R}^{2n+1} using essentially the argument above, together with the existence of a proper map $f: M \rightarrow \mathbb{R}$, given by Problem 2-30 (compare Guillemin and Pollack, *Differential Topology*). A much harder result of Whitney shows that M^n can actually be imbedded in \mathbb{R}^{2n} (H. Whitney, *The self-intersections of a smooth n -manifold in $2n$ -space*, Ann. of Math. 45 (1944), 220–246).

CHAPTER 4

TENSORS

All the constructions on vector bundles carried out in this chapter have a common feature. In each case, we replace each fibre $\pi^{-1}(p)$ by some other vector space, and then fit all these new vector spaces together to form a new vector bundle over the same base space.

The simplest case arises when we replace each fibre V by its dual space V^* . Recall that V^* denotes the vector space of all linear functions $\lambda: V \rightarrow \mathbb{R}$. If $f: V \rightarrow W$ is a linear transformation, then there is a linear transformation $f^*: W^* \rightarrow V^*$ defined by

$$(f^*\lambda)(v) = \lambda(fv).$$

It is clear that if $1_V: V \rightarrow V$ is the identity, then 1_V^* is the identity map of V^* and if $g: U \rightarrow V$, then $(f \circ g)^* = g^* \circ f^*$. These simple remarks already show that f^* is an isomorphism if $f: V \rightarrow W$ is, for $(f^{-1} \circ f)^* = 1_V^*$ and $(f \circ f^{-1})^* = 1_W^*$.

The dimension of V^* is the same as that of V , for finite dimensional V . In fact, if v_1, \dots, v_n is a basis for V , then the elements $v_i^* \in V^*$, defined by

$$v_i^*(v_j) = \delta_j^i,$$

are easily checked to be a basis for V^* . The linear function v_i^* depends on the entire set v_1, \dots, v_n , not just on v_i alone, and the isomorphism from V to V^* obtained by sending v_i to v_i^* is *not* independent of the choice of basis (consider what happens if v_1 is replaced by $2v_1$).

On the other hand, if $v \in V$, we can define $v^{**} \in V^{**} = (V^*)^*$ unambiguously by

$$v^{**}(\lambda) = \lambda(v) \quad \text{for every } \lambda \in V^*.$$

If $v^{**}(\lambda) = 0$ for every $\lambda \in V^*$, then $\lambda(v) = 0$ for all $\lambda \in V^*$, which implies that

$v = 0$. Thus the map $v \mapsto v^{**}$ is an isomorphism from V to V^{**} . It is called the **natural isomorphism** from V to V^{**} .

(Problem 6 gives a precise meaning to the word “natural”, formulated only after the term had long been in use. Once the meaning is made precise, we can prove that there is *no* natural isomorphism from V to V^* .)

Now let $\xi = \pi : E \rightarrow B$ be any vector bundle. Let

$$E' = \bigcup_{p \in B} [\pi^{-1}(p)]^*,$$

and define the function $\pi' : E' \rightarrow B$ to take each $[\pi^{-1}(p)]^*$ to p . If $U \subset B$, and $t : \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$ is a trivialization, then we can define a function

$$t' : \pi'^{-1}(U) \rightarrow U \times (\mathbb{R}^n)^*$$

in the obvious way: since the map t restricted to a fibre,

$$t_p : \pi^{-1}(p) \rightarrow \{p\} \times \mathbb{R}^n,$$

is an isomorphism, it gives us an isomorphism

$$(t_p^*)^{-1} : [\pi^{-1}(p)]^* \rightarrow \{p\} \times (\mathbb{R}^n)^*.$$

We can make $\pi' : E' \rightarrow B$ into a vector bundle, the **dual bundle** ξ^* of ξ , by requiring that all such t' be local trivializations. (We first pick an isomorphism from $(\mathbb{R}^n)^*$ to \mathbb{R}^n , once and for all.)

At first it might appear that $\xi^* \simeq \xi$, since each $\pi^{-1}(p)$ is isomorphic to $\pi'^{-1}(p)$. However, this is true merely because the two vector spaces have the same dimension. The lack of a natural isomorphism from V to V^* prevents us from constructing an equivalence between ξ^* and ξ . Actually, we will see later that in “most” cases ξ^* is equivalent to ξ ; for the present, readers may ponder this question for themselves. In contrast, the bundle $\xi^{**} = (\xi^*)^*$ is *always* equivalent to ξ . We construct the equivalence by mapping the fibre V of ξ over p to the fibre V^{**} of ξ^{**} over p by the natural isomorphism. If you

think about how ξ^* is constructed, it will appear obvious that this map is indeed an equivalence.

Even if ξ can be pictured geometrically (e.g., if ξ is TM), there is seldom a geometric picture for ξ^* . Rather, ξ^* operates on ξ : If s is a section of ξ and σ is a section of ξ^* , then we can define a function from B to \mathbb{R} by

$$p \mapsto \sigma(p)(s(p)) \quad \begin{array}{l} s(p) \in \pi^{-1}(p) \\ \sigma(p) \in \pi'^{-1}(p) = \pi^{-1}(p)^*. \end{array}$$

This function will be denoted simply by $\sigma(s)$.

When this construction is applied to the tangent bundle TM of M , the resulting bundle, denoted by T^*M , is called the **cotangent bundle** of M ; the fibre of T^*M over p is $(M_p)^*$. Like TM , the cotangent bundle T^*M is actually a C^∞ vector bundle: since two trivializations x_* and y_* of TM are C^∞ -related, the same is clearly true for x_*' and y_*' (in fact, $y_*' \circ (x_*')^{-1} = y_* \circ (x_*)^{-1}$). We can thus define C^∞ , as well as continuous, sections of T^*M . If ω is a C^∞ section of T^*M and X is a C^∞ vector field, then $\omega(X)$ is the C^∞ function $p \mapsto \omega(p)(X(p))$.

If $f: M \rightarrow \mathbb{R}$ is a C^∞ function, then a C^∞ section df of T^*M can be defined by

$$df(p)(X) = X(f) \quad \text{for } X \in M_p.$$

The section df is called the **differential** of f . Suppose, in particular, that X is $dc/dt|_{t_0}$, where $c(t_0) = p$. Recall that

$$\left. \frac{dc}{dt} \right|_{t_0} = c_* \left(\left. \frac{d}{dt} \right|_{t_0} \right).$$

This means that

$$\begin{aligned} df \left(\left. \frac{dc}{dt} \right|_{t_0} \right) &= c_* \left(\left. \frac{d}{dt} \right|_{t_0} \right) (f) \\ &= \left. \frac{d}{dt} \right|_{t_0} (f \circ c) \\ &= (f \circ c)'(t_0) \quad \text{or} \quad \left. \frac{d(f(c(t)))}{dt} \right|_{t_0}. \end{aligned}$$

Adopting the elliptical notations

$$\frac{dc}{dt} \quad \text{for} \quad \left. \frac{dc}{dt} \right|_t, \quad \frac{dg(t)}{dt} \quad \text{for} \quad g'(t),$$

this equation takes the nice form

$$df \left(\frac{dc}{dt} \right) = \frac{d(f(c(t)))}{dt}.$$

If (x, U) is a coordinate system, then the dx^i are sections of T^*M over U . Applying the definition, we see that

$$dx^i(p) \left(\left. \frac{\partial}{\partial x^j} \right|_p \right) = \delta_j^i.$$

Thus $dx^1(p), \dots, dx^n(p)$ is just the basis of M_p^* dual to the basis $\partial/\partial x^1|_p, \dots, \partial/\partial x^n|_p$ of M_p .

This means that every section ω can be expressed uniquely on U as

$$\omega(p) = \sum_{i=1}^n \omega_i(p) dx^i(p),$$

for certain functions ω_i on U . The section ω is continuous or C^∞ if and only if the functions ω_i are.

We can also write

$$\omega = \sum_{i=1}^n \omega_i dx^i,$$

if we define sums of sections and products of functions and sections in the obvious way (“pointwise” addition and multiplication).

The section df must have some such expression. In fact, we obtain a classical formula:

1. THEOREM. If (x, U) is a coordinate system and f is a C^∞ function, then on U we have

$$df = \sum_{i=1}^n \frac{\partial f}{\partial x^i} dx^i.$$

PROOF. If $X_p \in M_p$ is

$$X_p = \sum_{i=1}^n a^i \left. \frac{\partial}{\partial x^i} \right|_p,$$

then

$$a^i = X_p(x^i) = dx^i(p)(X_p).$$

Thus

$$\begin{aligned} df(p)(X_p) &= X_p(f) = \sum_{i=1}^n a^i \frac{\partial f}{\partial x^i}(p) \\ &= \sum_{i=1}^n \frac{\partial f}{\partial x^i}(p) dx^i(p)(X_p). \quad \spadesuit \end{aligned}$$

Classical differential geometers (and classical analysts) did not hesitate to talk about “infinitely small” changes dx^i of the coordinates x^i , just as Leibnitz had. No one wanted to admit that this was nonsense, because true results were obtained when these infinitely small quantities were divided into each other (provided one did it in the right way).

Eventually it was realized that the closest one can come to describing an infinitely small change is to describe a direction in which this change is supposed to occur, i.e., a tangent vector. Since df is supposed to be the infinitesimal change of f under an infinitesimal change of the point, df must be a function of this change, which means that df should be a function on tangent vectors. The dx^i themselves then metamorphosed into functions, and it became clear that they must be distinguished from the tangent vectors $\partial/\partial x^i$.

Once this realization came, it was only a matter of making new definitions, which preserved the *old* notation, and waiting for everybody to catch up. In short, all classical notions involving infinitely small quantities became functions on tangent vectors, like df , except for quotients of infinitely small quantities, which became tangent vectors, like dc/dt .

Looking back at the classical works from our modern vantage point, one can usually see that, no matter how obscurely expressed, this point of view was in

some sense the one always taken by classical geometers. In fact, the differential df was usually introduced in the following way:

CLASSICAL FORMULATION	MODERN FORMULATION
Let f be a function of the x^1, \dots, x^n , say $f = f(x^1, \dots, x^n)$.	Let f be a function on M , and x a coordinate system (so that $f = \tilde{f} \circ x$ for some function \tilde{f} on \mathbb{R}^n , namely $\tilde{f} = f \circ x^{-1}$).
Let x^i be functions of t , say $x^i = x^i(t)$. Then f becomes a function of t , $f(t) = f(x^1(t), \dots, x^n(t))$.	Let $c: \mathbb{R} \rightarrow M$ be a curve. Then $f \circ c: \mathbb{R} \rightarrow \mathbb{R}$, where $f \circ c(t) = \tilde{f}(x^1 \circ c(t), \dots, x^n \circ c(t))$.
<p>We now have</p> $\frac{df}{dt} = \sum_{i=1}^n \frac{\partial f}{\partial x^i} \frac{dx^i}{dt}.$ <p>(The classical notation, which suppresses the curve c, is still used by physicists, as we shall point out once again in Chapter 7.)</p>	<p>We now have</p> $\begin{aligned} (f \circ c)'(t) &= \sum_{i=1}^n D_i \tilde{f}(x(c(t))) \cdot (x^i \circ c)'(t) \\ &= \sum_{i=1}^n \frac{\partial f}{\partial x^i}(c(t)) \cdot (x^i \circ c)'(t) \end{aligned}$ <p>or</p> $\frac{d(f(c(t)))}{dt} = \sum_{i=1}^n \frac{\partial f}{\partial x^i}(c(t)) \cdot \frac{dx^i(c(t))}{dt}$
<p>Multiplying by dt gives</p> $df = \sum_{i=1}^n \frac{\partial f}{\partial x^i} dx^i.$ <p>(This equation signifies that true results are obtained by dividing by dt again, <i>no matter what the functions $x^i(t)$ are</i>. It is the closest approach in classical analysis to the realization of df as a function on tangent vectors.)</p>	<p>Consequently,</p> $df \left(\frac{dc}{dt} \right) = \sum_{i=1}^n \frac{\partial f}{\partial x^i}(c(t)) \cdot dx^i \left(\frac{dc}{dt} \right)$ <p>Since every tangent vector at $c(t)$ is of the form dc/dt, we have</p> $df = \sum_{i=1}^n \frac{\partial f}{\partial x^i} dx^i.$

In preparation for our reading of Gauss and Riemann, we will continually examine the classical way of expressing all concepts which we introduce. After a while, the “translation” of classical terminology becomes only a little more difficult than the translation of the German in which it was written.

Recall that if $f: M \rightarrow N$ is C^∞ , then there is a map $f_*: TM \rightarrow TN$; for each $p \in M$, we have a map $f_{*p}: M_p \rightarrow N_{f(p)}$. Since f_{*p} is a linear transformation between two vector spaces, it gives rise to a map

$$N_{f(p)}^* \rightarrow M_p^*.$$

Strict notational propriety would dictate that this map be denoted by $(f_{*p})^*$, but everyone denotes it simply by

$$f_p^*: N_{f(p)}^* \rightarrow M_p^*.$$

Notice that we cannot put all f_p^* together to obtain a bundle map from T^*N to T^*M ; in fact, the same $q \in N$ may be $f(p_i)$ for more than one $p_i \in M$, and there is no reason why f_{*p_1} should equal f_{*p_2} . On the other hand, we can do something with the cotangent bundle that we could not do with the tangent bundle. Suppose ω is a section of T^*N . Then we can define a section η of T^*M as follows:

$$\eta(p) = \omega(f(p)) \circ f_{*p},$$

i.e.,

$$\eta(p)(X_p) = \omega(f(p))(f_{*p}X_p) \quad \text{for } X_p \in M_p.$$

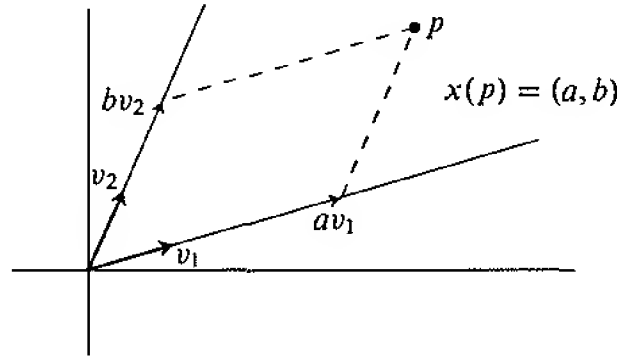
(The complex symbolism tends to hide the simple idea: to operate on a vector, we push it over to N by f_* , and then operate on it by ω .) This section η is denoted, naturally enough, by $f^*\omega$. There is no corresponding way of transferring a vector field X on M over to a vector field on N .

Despite these differences, we can say, roughly, that a map $f: M \rightarrow N$ produces a map f_* going in the same direction on the tangent bundle and a map f^* going in the opposite direction on the cotangent bundle. Nowadays such situations are always distinguished by calling the things which go in the same direction “covariant” and the things which go in the opposite direction “contravariant”. Classical terminology used these same words, and it just happens to have reversed this: a vector field is called a **contravariant vector field**, while a section of T^*M is called a **covariant vector field**. And no one has had the gall or authority to reverse terminology so sanctified by years of usage. So it’s very easy to remember which kind of vector field is covariant, and which contravariant—it’s just the opposite of what it logically ought to be.

The rationale behind the classical terminology can be seen by considering coordinate systems x on \mathbb{R}^n which are linear transformations. In this case, if $x(v_i) = e_i$, then

$$x(a^1 v_1 + \cdots + a^n v_n) = (a^1, \dots, a^n),$$

so the x coordinate system is just an “oblique Cartesian coordinate system”.



If x' is another such coordinate system, then $x'^j = \sum_{i=1}^n a_{ij} x^i$ for certain a_{ij} . Clearly $a_{ij} = \partial x'^j / \partial x^i$, so

$$(*) \quad x'^j = \sum_{i=1}^n \frac{\partial x'^j}{\partial x^i} x^i;$$

this can be seen directly from the fact that the matrix $(\partial x'^j / \partial x^i)$ is the constant matrix $D(x' \circ x^{-1}) = x' \circ x^{-1}$. Comparing $(*)$ with

$$(**) \quad dx'^j = \sum_{i=1}^n \frac{\partial x'^j}{\partial x^i} dx^i,$$

from Theorem 1, we see that the differentials dx^i “change in the same way” as the coordinates x^i , hence they are “covariant”. Consequently, any combination

$$\omega = \sum_{i=1}^n \omega_i dx^i$$

is also called “covariant”. Notice that if we also have

$$\omega = \sum_{i=1}^n \omega'_i dx'^i,$$

then we can express the ω'_i in terms of the ω_i . Substituting

$$dx^i = \sum_{j=1}^n \frac{\partial x^i}{\partial x'^j} dx'^j$$

into the first expression for ω and comparing coefficients with the second, we find that

$$\omega'_j = \sum_{i=1}^n \omega_i \frac{\partial x^i}{\partial x'^j}.$$

On the other hand, given two expressions

$$\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} = \sum_{i=1}^n a'^i \frac{\partial}{\partial x'^i}$$

for a vector field, the functions a'^j must satisfy

$$a'^j = \sum_{i=1}^n a^i \frac{\partial x'^j}{\partial x^i}.$$

These expressions can always be remembered by noting that indices which are summed over always appear once “above” and once “below”. (Coordinate functions x^1, \dots, x^n used to be denoted by x_1, \dots, x_n . This suggested subscripts ω_i for covariant vector fields and superscripts a^i for contravariant vector fields. After this was firmly established, the indices on the x ’s were shifted upstairs again to make the summation convention work out.)

Covariant and contravariant vector fields, i.e., sections of T^*M and TM , respectively, are also called covariant and contravariant tensors (or tensor fields) of order 1, which is a warning that worse things are to come. We begin with some worse algebra.

If V_1, \dots, V_m are vector spaces, a function

$$T: V_1 \times \dots \times V_m \rightarrow \mathbb{R}$$

is **multilinear** if

$$v \mapsto T(v_1, \dots, v_{k-1}, v, v_{k+1}, \dots, v_m)$$

is linear for each choice of $v_1, \dots, v_{k-1}, v_{k+1}, \dots, v_m$. The set of all such T is clearly a vector space. If $V_1, \dots, V_m = V$, this vector space will be denoted

by $\mathcal{T}^m(V)$. Notice that $\mathcal{T}^1(V) = V^*$. If $f: V \rightarrow W$ is a linear transformation, then there is a linear transformation $f^*: \mathcal{T}^m(W) \rightarrow \mathcal{T}^m(V)$, defined completely analogously to the case $m = 1$:

$$f^*T(v_1, \dots, v_m) = T(f(v_1), \dots, f(v_m)).$$

For $T \in \mathcal{T}^k(V)$, and $S \in \mathcal{T}^l(V)$ we can define the “tensor product” $T \otimes S \in \mathcal{T}^{k+l}(V)$ by

$$T \otimes S(v_1, \dots, v_k, v_{k+1}, \dots, v_{k+l}) = T(v_1, \dots, v_k) \cdot S(v_{k+1}, \dots, v_{k+l}).$$

Of course, $T \otimes S$ is not $S \otimes T$. On the other hand, $(S \otimes T) \otimes U = S \otimes (T \otimes U)$, so we can define n -fold tensor products unambiguously; this tensor product operation is itself multilinear, $(S_1 + S_2) \otimes T = S_1 \otimes T + S_2 \otimes T$, etc. In particular, if v_1, \dots, v_n is a basis for V and v_1^*, \dots, v_n^* is the dual basis for $V^* = \mathcal{T}^1(V)$, then the elements

$$v_{i_1}^* \otimes \cdots \otimes v_{i_k}^* \quad 1 \leq i_1, \dots, i_k \leq n$$

are easily seen to be a basis for $\mathcal{T}^k(V)$, which thus has dimension n^k .

We can use this new algebraic construction to obtain a new bundle from any vector bundle $\xi = \pi: E \rightarrow B$. We let

$$E' = \bigcup_{p \in B} \mathcal{T}^k(\pi^{-1}(p)),$$

and let

$$\pi': E' \rightarrow B \quad \text{take} \quad \mathcal{T}^k(\pi^{-1}(p)) \quad \text{to} \quad p.$$

If $U \subset B$ and

$$t: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$$

is a trivialization, then the isomorphisms

$$t_p: \pi^{-1}(p) \rightarrow \{p\} \times \mathbb{R}^n$$

yield isomorphisms

$$(t_p^*)^{-1}: \mathcal{T}^k(\pi^{-1}(p)) \rightarrow \{p\} \times \mathcal{T}^k(\mathbb{R}^n).$$

If we choose an isomorphism $\mathcal{T}^k(\mathbb{R}^n) \rightarrow \mathbb{R}^{n^k}$ once and for all, these maps can be put together to give a map

$$t': \pi'^{-1}(U) \rightarrow U \times \mathbb{R}^{n^k}.$$

We make $\pi': E' \rightarrow B$ into a vector bundle $\mathcal{T}^k(\xi)$ by requiring that all such ι' be local trivializations. The bundle ξ^* is the special case $k = 1$.

For the case of TM , the bundle $\mathcal{T}^k(TM)$ is called the **bundle of covariant tensors of order k** , and a section is called a **covariant tensor field of order k** . If (x, U) is a coordinate system, so that

$$dx^1(p), \dots, dx^n(p)$$

is a basis for $(M_p)^*$, then the k -fold tensor products

$$dx^{i_1}(p) \otimes \dots \otimes dx^{i_k}(p) \in \mathcal{T}^k(M_p) \quad 1 \leq i_1, \dots, i_k \leq n$$

are a basis for $\mathcal{T}^k(M_p)$. Thus, on U every covariant tensor field A of order k can be written

$$A(p) = \sum_{i_1, \dots, i_k} A_{i_1 \dots i_k}(p) dx^{i_1}(p) \otimes \dots \otimes dx^{i_k}(p),$$

or simply

$$A = \sum_{i_1, \dots, i_k} A_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k},$$

where $dx^{i_1} \otimes \dots \otimes dx^{i_k}$ now denotes a section of $\mathcal{T}^k(TM)$. If we also have

$$A = \sum_{i_1, \dots, i_k} A'_{i_1 \dots i_k} dx'^{i_1} \otimes \dots \otimes dx'^{i_k},$$

then

$$A'_{\alpha_1 \dots \alpha_k} = \sum_{i_1, \dots, i_k} A_{i_1 \dots i_k} \frac{\partial x^{i_1}}{\partial x'^{\alpha_1}} \dots \frac{\partial x^{i_k}}{\partial x'^{\alpha_k}}$$

(the products are just ordinary products of functions). To derive this equation, we just use equation (**) on page 114, and multilinearity of \otimes . The section A is continuous or C^∞ if and only if the functions $A_{i_1 \dots i_k}$ are.

A covariant tensor field A of order k can just be thought of as an operation \bar{A} on k vector fields X_1, \dots, X_k which yields a function:

$$\bar{A}(X_1, \dots, X_k)(p) = A(p)(X_1(p), \dots, X_k(p)).$$

Notice that \bar{A} is multilinear on the set \mathcal{V} of C^∞ vector fields:

$$\bar{A}(X_1, \dots, X_i + X'_i, \dots, X_k) = \bar{A}(X_1, \dots, X_i, \dots, X_k) + \bar{A}(X_1, \dots, X'_i, \dots, X_k)$$

$$\bar{A}(X_1, \dots, aX_i, \dots, X_k) = a\bar{A}(X_1, \dots, X_k).$$

Moreover, because \bar{A} is defined “pointwise”, it is actually linear over the C^∞ functions \mathcal{F} ; i.e., if f is C^∞ , then

$$\bar{A}(X_1, \dots, fX_i, \dots, X_k) = f\bar{A}(X_1, \dots, X_i, \dots, X_k),$$

for we have

$$\begin{aligned}\bar{A}(X_1, \dots, fX_i, \dots, X_k)(p) &= A(p)(X_1(p), \dots, f(p)X_i(p), \dots, X_k(p)) \\ &= f(p)A(p)(X_1(p), \dots, X_i(p), \dots, X_k(p)) \\ &= f(p) \cdot \bar{A}(X_1, \dots, X_i, \dots, X_k)(p).\end{aligned}$$

We are finally ready for another theorem, one that is used over and over.

2. THEOREM. If

$$\mathcal{A}: \underbrace{\mathcal{V} \times \dots \times \mathcal{V}}_{k \text{ times}} \rightarrow \mathcal{F}$$

is linear over \mathcal{F} , then there is a unique tensor field A with $\mathcal{A} = \bar{A}$.

PROOF. Note first that if $v \in M_p$ is any tangent vector, then there is a vector field $X \in \mathcal{V}$ with $X(p) = v$. In fact, if (x, U) is a coordinate system and

$$v = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p,$$

then we can define

$$X = \begin{cases} f \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} & \text{on } U \\ 0 & \text{outside } U, \end{cases}$$

where each a^i now denotes a constant function and f is a C^∞ function with $f(p) = 1$ and support $f \subset U$.

Now if $v_1, \dots, v_k \in M_p$ are extended to vector fields $X_1, \dots, X_k \in \mathcal{V}$ we clearly must define

$$A(p)(v_1, \dots, v_k) = \mathcal{A}(X_1, \dots, X_k)(p).$$

The problem is to prove that this is well-defined: If $X_i(p) = Y_i(p)$ for each i , we claim that

$$\mathcal{A}(X_1, \dots, X_k)(p) = \mathcal{A}(Y_1, \dots, Y_k)(p).$$

(The map \mathcal{A} “lives at points”, to use the in terminology.) For simplicity, take the case $k = 1$ (the general case is exactly analogous). The proof that $\mathcal{A}(X)(p) = \mathcal{A}(Y)(p)$ when $X(p) = Y(p)$ is in two steps.

(1) Suppose first that $X = Y$ in a neighborhood U of p . Let f be a C^∞ function with $f(p) = 1$ and support $f \subset U$. Then $fX = fY$, so

$$f\mathcal{A}(X) = \mathcal{A}(fX) = \mathcal{A}(fY) = f\mathcal{A}(Y);$$

evaluating at p gives

$$\mathcal{A}(X)(p) = \mathcal{A}(Y)(p).$$

(2) To prove the result, it obviously suffices to show that $\mathcal{A}(X)(p) = 0$ if $X(p) = 0$. Let (x, U) be a coordinate system around p , so that on U we can write

$$X = \sum_{i=1}^n b^i \frac{\partial}{\partial x^i} \quad \text{where all } b^i(p) = 0.$$

If g is 1 in a neighborhood V of p , and support $g \subset U$, then

$$Y = g \sum_{i=1}^n b^i \frac{\partial}{\partial x^i} = \sum_{i=1}^n b^i g \frac{\partial}{\partial x^i}$$

is a well-defined C^∞ vector field on all of M which equals X on V , so that

$$\mathcal{A}(X)(p) = \mathcal{A}(Y)(p), \quad \text{by (1).}$$

Now

$$\begin{aligned} \mathcal{A}(Y)(p) &= \sum_{i=1}^n b^i(p) \cdot \mathcal{A}\left(g \frac{\partial}{\partial x^i}\right)(p) \\ &= 0, \quad \text{since } b^i(p) = 0. \quad \spadesuit \end{aligned}$$

Because of Theorem 2, we will never distinguish between the tensor field A and the operation \bar{A} , nor will we use the symbol \bar{A} any longer. Note that Theorem 2 applies, in particular, to the case $k = 1$, where $\mathcal{T}^k(TM) = T^*M$, the cotangent bundle: a function from $\mathcal{V} \rightarrow \mathcal{F}$ which is linear over \mathcal{F} comes from a covariant vector field ω . Just as with covariant vector fields, a C^∞ map $f: M \rightarrow N$ gives a map f^* taking covariant tensor fields A of order k on N to covariant tensor fields f^*A of order k on M :

$$f^*A(p)(X_{1p}, \dots, X_{kp}) = A(f(p))(f_*X_{1p}, \dots, f_*X_{kp}).$$

Moreover, if A and B are covariant tensor fields of orders k and l , respectively, then we can define a new covariant tensor field $A \otimes B$ of order $k + l$:

$$(A \otimes B)(p) = A(p) \otimes B(p) \quad (\text{operating on } M_p \times \cdots \times M_p \text{ } k + l \text{ times}).$$

Although covariant tensor fields will be our main concern, if only for the sake of completeness we should define contravariant tensor fields. Recall that a contravariant vector field is a section X of TM . So each $X_p \in M_p$. Now an element v of a vector space V can be thought of as a linear function $v: V^* \rightarrow \mathbb{R}$; we just define $v(\lambda)$ to be $\lambda(v)$. A contravariant tensor field of order k is just a section A of the bundle $\mathcal{T}^k(T^*M)$; thus, each $A(p)$ is a k -linear function on M_p^* . We could also use the notation $\mathcal{T}_k(TM)$, if we use $\mathcal{T}_k(V)$ to denote all k -linear functions on V^* . In local coordinates we can write

$$A(p) = \sum_{j_1, \dots, j_k} A^{j_1 \dots j_k}(p) \left. \frac{\partial}{\partial x^{j_1}} \right|_p \otimes \cdots \otimes \left. \frac{\partial}{\partial x^{j_k}} \right|_p$$

(remember that each $\partial/\partial x^j|_p$ operates on M_p^*), or simply

$$A = \sum_{j_1, \dots, j_k} A^{j_1 \dots j_k} \frac{\partial}{\partial x^{j_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{j_k}}.$$

If we have another such expression,

$$A = \sum_{j_1, \dots, j_k} A'^{j_1 \dots j_k} \frac{\partial}{\partial x'^{j_1}} \otimes \cdots \otimes \frac{\partial}{\partial x'^{j_k}},$$

then we easily compute that

$$A'^{\beta_1 \dots \beta_k} = \sum_{j_1, \dots, j_k} A^{j_1 \dots j_k} \frac{\partial x'^{\beta_1}}{\partial x^{j_1}} \cdots \frac{\partial x'^{\beta_k}}{\partial x^{j_k}}.$$

A contravariant tensor field A of order k can be considered as an operator \bar{A} taking k covariant vector fields $\omega_1, \dots, \omega_k$ into a function:

$$\bar{A}(\omega_1, \dots, \omega_k)(p) = A(p)(\omega_1(p), \dots, \omega_k(p)).$$

Naturally, there is an analogue of Theorem 2, proved exactly the same way, that allows us to dispense with the notation \bar{A} , and to identify contravariant tensor fields of order k with operators on k covariant vector fields that are *linear over the C^∞ functions \mathcal{F}* .

Finally, we are ready to introduce “mixed” tensor fields. To make the introduction less painful, we consider a special case first. If V is a vector space, let $\mathcal{T}_1^1(V)$ denote all bilinear functions

$$T: V \times V^* \rightarrow \mathbb{R}.$$

A vector bundle $\xi = \pi: E \rightarrow B$ gives rise to a vector bundle $\mathcal{T}_1^1(\xi)$, obtained by replacing each fibre $\pi^{-1}(p)$ by $\mathcal{T}_1^1(\pi^{-1}(p))$. In particular, sections of $\mathcal{T}_1^1(TM)$ are called tensor fields, covariant of order 1 and contravariant of order 1.

There are all sorts of algebraic tricks one can play with $\mathcal{T}_1^1(V)$; although they should be kept to a minimum, certain ones are quite important. Let $End(V)$ denote the vector space of all linear transformations $T: V \rightarrow V$ (“endomorphisms” of V). Notice that each $S \in End(V)$ gives rise to a bilinear $\bar{S} \in \mathcal{T}_1^1(V)$,

$$\bar{S}: V \times V^* \rightarrow \mathbb{R},$$

by the formula

$$(*) \quad \bar{S}(v, \lambda) = \lambda(S(v)).$$

Moreover, the correspondence $S \mapsto \bar{S}$ from $End(V)$ to $\mathcal{T}_1^1(V)$ is linear and one-one, for $\bar{S} = 0$ implies that $\lambda(S(v)) = 0$ for all λ , which implies that $S(v) = 0$, for all v . Since both $End(V)$ and $\mathcal{T}_1^1(V)$ have dimension n^2 , this map is an isomorphism. The inverse, however, is not so easy to describe. Given \bar{S} , for each v the vector $S(v) \in V$ is merely determined by describing the action of a λ on it according to (*). It is not hard to check that this isomorphism of $End(V)$ and $\mathcal{T}_1^1(V)$ makes the identity map $1: V \rightarrow V$ in $End(V)$ correspond to the “evaluation” map

$$e: V \times V^* \rightarrow \mathbb{R} \quad \text{in } \mathcal{T}_1^1(V)$$

given by

$$e(v, \lambda) = \lambda(v).$$

Generally speaking, our isomorphism can be used to transfer any operation from $End(V)$ to $\mathcal{T}_1^1(V)$. In particular, given a bilinear

$$T: V \times V^* \rightarrow \mathbb{R},$$

we can take the trace of the corresponding $S: V \rightarrow V$; this number is called the contraction of T . If v_1, \dots, v_n is a basis of V and

$$T = \sum_{i,j} T_i^j v^*_i \otimes v_j,$$

then we can find the matrix $A = (a_{ij})$ of S , defined by

$$S(v_i) = \sum_{j=1}^n a_{ji} v_j,$$

in terms of the T_i^j ; in fact,

$$a_{ji} = v_j^*(S(v_i)) = T(v_i, v_j^*) = T_i^j.$$

Thus

$$\text{contraction of } T = \sum_{i=1}^n T_i^i.$$

(The term “contraction” comes from the fact that the number of indices is contracted from 2 to 0 by setting the upper and lower indices equal and summing.)

These identifications and operations can be carried out, fibre by fibre, in any fibre bundle $\mathcal{T}_1^1(\xi)$. Thus, a section A of $\mathcal{T}_1^1(\xi)$ can just as well be considered as a section of the bundle $\text{End}(\xi)$, obtained by replacing each fibre $\pi^{-1}(p)$ by $\text{End}(\pi^{-1}(p))$. In this case, each $A(p)$ is an endomorphism of $\pi^{-1}(p)$. Moreover, each section A gives rise to a *function*

$$(\text{contraction of } A): B \rightarrow \mathbb{R}$$

defined by

$$\begin{aligned} p &\mapsto \text{contraction of } A(p) \\ &= \text{trace } A(p) \quad \text{if we consider } A(p) \in \text{End}(\pi^{-1}(p)). \end{aligned}$$

In particular, given a tensor field A , covariant of order 1 and contravariant of order 1, which is a section of $\mathcal{T}_1^1(TM)$, we can consider each $A(p)$ as an endomorphism of M_p , and we obtain a function “contraction of A ”. If in a coordinate system

$$A = \sum_{i,j} A_i^j dx^i \otimes \frac{\partial}{\partial x^j},$$

then

$$(\text{contraction of } A) = \sum_{i=1}^n A_i^i.$$

The general notion of a mixed tensor field is a straightforward generalization. Define $\mathcal{T}_l^k(V)$ to be the set of all $(k + l)$ -linear

$$T: \underbrace{V \otimes \cdots \otimes V}_{k \text{ times}} \times \underbrace{V^* \otimes \cdots \otimes V^*}_{l \text{ times}} \rightarrow \mathbb{R}.$$

Every bundle ξ gives rise to a bundle $\mathcal{T}_i^k(\xi)$. Sections of $\mathcal{T}_i^k(TM)$ are called tensor fields, covariant of order k and contravariant of order l , or simply of type $\binom{k}{l}$, an abbreviation that also saves everybody embarrassment about the use of the words “covariant” and “contravariant”. Locally, a tensor field A of type $\binom{k}{l}$ can be expressed as

$$A = \sum_{\substack{i_1, \dots, i_k \\ j_1, \dots, j_l}} A_{i_1 \dots i_k}^{j_1 \dots j_l} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}},$$

and if

$$A' = \sum_{\substack{i_1, \dots, i_k \\ j_1, \dots, j_l}} A'_{i_1 \dots i_k}^{j_1 \dots j_l} dx'^{i_1} \otimes \dots \otimes dx'^{i_k} \otimes \frac{\partial}{\partial x'^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x'^{j_l}},$$

then

$$(*) \quad A'_{\alpha_1 \dots \alpha_k}^{\beta_1 \dots \beta_l} = \sum_{\substack{i_1, \dots, i_k \\ j_1, \dots, j_l}} A_{i_1 \dots i_k}^{j_1 \dots j_l} \frac{\partial x^{i_1}}{\partial x'^{\alpha_1}} \dots \frac{\partial x^{i_k}}{\partial x'^{\alpha_k}} \frac{\partial x'^{\beta_1}}{\partial x^{j_1}} \dots \frac{\partial x'^{\beta_l}}{\partial x^{j_l}}.$$

Classical differential geometry books are filled with monstrosities like this equation. In fact, the classical definition of a tensor field is: an assignment of n^{k+l} functions to every coordinate system so that $(*)$ holds between the n^{k+l} functions assigned to any two coordinate systems x and x' . (!) Or even, “a set of n^{k+l} functions which changes according to $(*)$ ”. Consequently, in classical differential geometry, all important tensors are actually defined by defining the functions A_i^j , in terms of the coordinate system x , and then checking that $(*)$ holds.

Here is an important example. In every classical differential geometry book, one will find the following assertion: “The Kronecker delta δ_i^j is a tensor.” In other words, it is asserted that if one chooses the same n^2 functions δ_i^j for each coordinate system, then $(*)$ holds, i.e.,

$$\delta_\alpha^\beta = \sum_{i,j} \delta_i^j \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x'^\beta}{\partial x^j};$$

this is certainly true, for

$$\sum_{i,j} \delta_i^j \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x'^\beta}{\partial x^j} = \sum_{i=1}^n \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x'^\beta}{\partial x^i} = \delta_\alpha^\beta.$$

From our point of view, what this equation shows is that

$$A = \sum_{i,j} \delta_i^j dx^i \otimes \frac{\partial}{\partial x^j}$$

is a certain tensor field, independent of the choice of the coordinate system x . To identify the mysterious map

$$A(p): M_p \times M_p^* \rightarrow \mathbb{R},$$

we consider $v \in M_p$ and $\lambda \in M_p^*$ with the expressions

$$v = \sum_{\alpha=1}^n a^\alpha \frac{\partial}{\partial x^\alpha} \Big|_p, \quad \lambda = \sum_{\beta=1}^n b_\beta dx^\beta(p);$$

then

$$\begin{aligned} A(p)(v, \lambda) &= \sum_{i,j} \delta_i^j dx^i(p) \otimes \frac{\partial}{\partial x^j} \Big|_p (v, \lambda) \\ &= \sum_{i,j} \delta_i^j dx^i(p) \left(\sum_{\alpha=1}^n a^\alpha \frac{\partial}{\partial x^\alpha} \Big|_p \right) \cdot \frac{\partial}{\partial x^j} \Big|_p \left(\sum_{\beta=1}^n b_\beta dx^\beta(p) \right) \\ &= \sum_{i,j} \delta_i^j a^i b_j \\ &= \sum_{i=1}^n a^i b_i \\ &= \lambda(v). \end{aligned}$$

Thus $A(p)$ is just the evaluation map $M_p \times M_p^* \rightarrow \mathbb{R}$; considered as an endomorphism of M_p , it is just the identity map.

The contraction of a tensor is defined, classically, in a similar manner. Given a tensor, i.e., a collection of functions A_i^j , one for each coordinate system, satisfying

$$A'^\beta_\alpha = \sum_{i,j} A_i^j \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x'^\beta}{\partial x^j},$$

we note that

$$\begin{aligned}
 \sum_{\alpha=1}^n A'^{\alpha}_{\alpha} &= \sum_{\alpha=1}^n \left(\sum_{i,j} A_i^j \frac{\partial x^i}{\partial x'^{\alpha}} \frac{\partial x'^{\alpha}}{\partial x^j} \right) \\
 &= \sum_{i,j} A_i^j \sum_{\alpha=1}^n \frac{\partial x^i}{\partial x'^{\alpha}} \frac{\partial x'^{\alpha}}{\partial x^j} \\
 &= \sum_{i,j} A_i^j \delta_j^i \\
 &= \sum_{i=1}^n A_i^i,
 \end{aligned}$$

so that this sum is a well-defined function. This calculation tends to obscure the one part which is really necessary—verification of the fact that the trace of a linear transformation, defined as the sum of the diagonal entries of its matrix, is independent of the basis with respect to which the matrix is written.

Incidentally, a tensor of type $\binom{k}{l}$ can be contracted with respect to any pair of upper and lower indices. For example, the functions

$$B_{\mu\nu}^{\beta\gamma} = \sum_{\alpha=1}^n A_{\mu\nu\alpha}^{\alpha\beta\gamma}$$

“transform correctly” if the $A_{\mu\nu\lambda}^{\alpha\beta\gamma}$ do. If we consider each $A(p) \in \mathcal{T}_3^3(M_p)$, then we are taking $B(p) \in \mathcal{T}_2^2(M_p)$ to be

$$B(p)(v_1, v_2, \lambda_1, \lambda_2) = \text{contraction of: } (v, \lambda) \mapsto A(p)(v, v_1, v_2, \lambda_1, \lambda_2, \lambda).$$

While a contravariant vector field is classically a set of n functions which “transforms in a certain way”, a vector at a single point p is classically just an assignment of n numbers a^1, \dots, a^n to each coordinate system x , such that the numbers a'^1, \dots, a'^n assigned to x' satisfy

$$a'^j = \sum_{i=1}^n a^i \frac{\partial x'^j}{\partial x^i}(p).$$

This is *precisely* the definition we adopted when we defined tangent vectors as equivalence classes $[x, a]_p$. The revolution in the modern approach is that the set of all vectors is made into a bundle, so that vector fields can be defined as sections, rather than as equivalence classes of sets of functions, and that all other

types of tensors are constructed from this bundle. The tangent bundle itself was almost a victim of the excesses of revolutionary zeal. For a long time, the party line held that TM must be defined either as derivations, or as equivalence classes of curves; the return to the old definition was influenced by the “functorial” point of view of Theorems 3-1 and 3-4.

The modern revolt against the classical point of view has been so complete in certain quarters that some mathematicians will give a three page proof that avoids coordinates in preference to a three line proof that uses them. We won’t go quite that far, but we will give an “invariant” definition (one that does not use a coordinate system) of any tensors that are defined. Unlike the “Kronecker delta” and contractions, such invariant definitions are usually not so easy to come by. As we shall see, invariant definitions of all the important tensors in differential geometry are made by means of Theorem 2. We seldom define $A(p)$ directly; instead we define a function \mathcal{A} on vector fields, which miraculously turns out to be linear over the C^∞ functions \mathcal{F} , and hence must come from some A . At the appropriate time we will discuss whether or not this is all a big cheat.

PROBLEMS

1. Let $f: M^n \rightarrow N^m$, and suppose that (x, U) and (y, V) are coordinate systems around p and $f(p)$, respectively.

(a) If $g: N \rightarrow \mathbb{R}$, then

$$\frac{\partial(g \circ f)}{\partial x^i}(p) = \sum_{j=1}^m \frac{\partial g}{\partial y^j}(f(p)) \cdot \frac{\partial(y^j \circ f)}{\partial x^i}(p).$$

(Proposition 2-3 is the special case $f = \text{identity}$.)

(b) Show that

$$f_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) = \sum_{j=1}^m \frac{\partial(y^j \circ f)}{\partial x^i}(p) \cdot \frac{\partial}{\partial y^j} \Big|_{f(p)},$$

and, more generally, express $f_* \left(\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p \right)$ in terms of the $\partial/\partial y^j|_{f(p)}$.

(c) Show that

$$(f^* dy^j)(p) = \sum_{i=1}^n \frac{\partial(y^j \circ f)}{\partial x^i}(p) \cdot dx^i(p).$$

(d) Express

$$f^* \left(\sum_{j_1, \dots, j_n} a_{j_1 \dots j_n} dy^{j_1} \otimes \dots \otimes dy^{j_n} \right)$$

in terms of the dx^i .

2. If $f, g: M \rightarrow N$ are C^∞ , show that

$$d(fg) = f dg + g df.$$

3. Let $f: M \rightarrow \mathbb{R}$ be C^∞ . For $v \in M_p$, show that

$$f_*(v) = df(v)_{f(p)} \in \mathbb{R}_{f(p)}.$$

4. (a) Show that if the ordered bases v_1, \dots, v_n and w_1, \dots, w_n for V are equally oriented, then the same is true of the bases v_1^*, \dots, v_n^* and w_1^*, \dots, w_n^* for V^* .

(b) Show that a bundle ξ is orientable if and only if ξ^* is orientable.

5. The following statements and problems are all taken from Eisenhart's classical work *Riemannian Geometry*. In each case, check them, using the classical methods, and then translate the problem and solution into modern terms. An "invariant" is just a (well-defined) function. Remember that the summation convention is always used, so $\lambda^i \mu_i$ means $\sum_{i=1}^n \lambda^i \mu_i$. Hints and answers are given at the end, after (xiii).

(i) If the quantity $\lambda^i \mu_i$ is an invariant and either λ^i or μ_i are the components of an arbitrary [covariant or contravariant] vector field, the other sets are components of a vector field.

(ii) If λ_{α}^i are the components of n vector fields [in an n -manifold], where i for $i = 1, \dots, n$ indicates the component and α for $\alpha = 1, \dots, n$ the vector, and these vectors are independent, that is, $\det(\lambda_{\alpha}^i) \neq 0$, then any vector-field λ^i is expressible in the form

$$\lambda^i = a^{\alpha} \lambda_{\alpha}^i,$$

where the a 's are invariants.

(iii) If μ_i are the components of a given vector-field, any vector-field λ^i satisfying $\lambda^i \mu_i = 0$ is expressible linearly in terms of $n - 1$ independent vector fields λ_{α}^i for $\alpha = 1, \dots, n - 1$ which satisfy the equation.

(iv) If $a^{ij} = a^{ji}$ for the components of a tensor field in one coordinate system, then $a'^{ij} = a'^{ji}$ for the coordinates in any other coordinate system.

(v) If a^{ij} and b^{ij} are components of a tensor field, so are $a^{ij} + b^{ij}$. If a^{ij} and b_{kl} are components of a tensor field, so are $a^{ij} b_{kl}$.

(vi) If $a_{ij} \lambda^i \lambda^j$ is an invariant for λ^i an arbitrary vector, then $a_{ij} + a_{ji}$ are the components of a tensor; in particular, if $a_{ij} \lambda^i \lambda^j = 0$, then $a_{ij} + a_{ji} = 0$.

(vii) If $a_{ij} \lambda^i \lambda^j = 0$ for all vectors λ^i such that $\lambda^i \mu_i = 0$, where μ_i is a given covariant vector, if v^i is defined [c.f. (iii)] by $a_{ij} \lambda_{\alpha}^i v^j = 0$, $\alpha = 1, \dots, n - 1$ and $\mu_i v^i \neq 0$, and by definition

$$a_{ij} v^i = \sigma_j \quad v^i \mu_i = \tau,$$

then $(a_{ij} - \frac{1}{\tau} \mu_i \sigma_j) \xi^i \xi^j = 0$ is satisfied by every vector field ξ^i , and consequently

$$a_{ij} + a_{ji} = \frac{1}{\tau} (\mu_i \sigma_j + \mu_j \sigma_i).$$

(viii) If a_{rs} are the components of a tensor and b and c are invariants, show that if $ba_{rs} + ca_{sr} = 0$, then either $b = -c$ and a_{rs} is symmetric, or $b = c$ and a_{rs} is skew-symmetric.

(ix) By definition the *rank* of a tensor of the second order a_{ij} is the rank of the matrix (a_{ij}) . Show that the rank is invariant under all transformations of coordinates.

(x) Show that the rank of the tensor of components $a_i b_j$, where a_i and b_j are the components of two vectors, is one; show that for the symmetric tensor $a_i b_j + a_j b_i$ the rank is two.

(xi) Show that the tensor equation $a^i_j \lambda_i = \alpha \lambda_j$, where α is an invariant, can be written in the form $(a^i_j - \alpha \delta^i_j) \lambda_i = 0$. Show also that $a^i_j = \delta^i_j \alpha$, if the equation is to hold for an arbitrary vector λ_i .

(xii) If $a^i_j \lambda_i = \alpha \lambda_j$ holds for all vectors λ_i such that $\mu^i \lambda_i = 0$, where μ^i is a given vector, then

$$a^i_j = \alpha \delta^i_j + \sigma_j \mu^i.$$

(xiii) If

$$\delta_{i_1 \dots i_p}^{j_1 \dots j_p} = \begin{cases} 0 & \text{if } j_\alpha = j_\beta \text{ for some } \alpha \neq \beta \text{ or } i_\alpha = i_\beta \text{ for some } \alpha \neq \beta \\ & \text{or if } \{j_1, \dots, j_p\} \neq \{i_1, \dots, i_p\} \\ 1 & \text{if } j_1, \dots, j_p \text{ is an even permutation of } i_1, \dots, i_p \\ -1 & \text{if } j_1, \dots, j_p \text{ is an odd permutation of } i_1, \dots, i_p \end{cases}$$

then $\delta_{i_1 \dots i_p}^{j_1 \dots j_p}$ are the components of a tensor in all coordinate systems.

HINTS AND ANSWERS.

(i) ω is determined if $\omega(X)$ is known for all X , and *vice versa*.

(iii) Given ω [with $\omega(p) \neq 0$ for all p], there are everywhere linearly independent vector fields X_1, \dots, X_{n-1} which span $\ker \omega$ at each point. (This is true only locally. For example, on $S^2 \times \mathbb{R}$ there is an ω such that $\ker \omega(p, t)$ consists of vectors tangent to $S^2 \times \{t\}$.)

(vi) For $T: V \times V \rightarrow \mathbb{R}$, let $T'(v, w) = T(w, v)$. Then $T + T'$ is determined by $S(v) = T(v, v)$. For, $T(v + w, v + w) = T(v, v) + T(v, w) + T(w, v) + T(w, w)$. Similarly, $T(v, v) = 0$ for all v implies that $T + T' = 0$.

(vii) Given ω [with $\omega(p) \neq 0$ for all p], choose Y complementary to $\ker \omega$ at all points. If $\sigma(Z) = T(Y, Z)$, then $T(Z, Z) = \omega(Z)\sigma(Z)/\omega(Y)$ for all vector fields Z .

(ix) $T: V \times V \rightarrow \mathbb{R}$ corresponds to $\bar{T}: V \rightarrow V^*$ [where $\bar{T}(v)(w) = T(v, w)$]. The rank of T may be defined as the rank of \bar{T} (consider the matrix of \bar{T} with respect to bases v_1, \dots, v_n and v^*_1, \dots, v^*_n).

(xii) Let $V = M_p^*$. If $T: V \rightarrow V$ and $\mu \in V^*$ and $T(v) = \alpha v$ for all $v \in \ker \mu$, there is a y complementary to $\ker \mu$ such that

$$T(v) = \alpha v + \mu(v)y \quad \text{for all } v.$$

(Begin by choosing y_0 complementary to $\ker \mu$ and writing v uniquely as $v_0 + cy_0$ for $v_0 \in \ker \mu$.)

(xiii) Define

$$\delta: \underbrace{V \times \cdots \times V}_p \times \underbrace{V^* \times \cdots \times V^*}_p \rightarrow \mathbb{R}$$

by

$$\delta(v_1, \dots, v_p, \lambda_1, \dots, \lambda_p) = \det(\lambda_i(v_j)).$$

6. (a) Let $i_V: V \rightarrow V^{**}$ be the “natural isomorphism” $i_V(v)(\lambda) = \lambda(v)$. Show that for any linear transformation $f: V \rightarrow W$, the following diagram commutes:

$$\begin{array}{ccc} V & \xrightarrow{i_V} & V^{**} \\ f \downarrow & & \downarrow f^{**} \\ W & \xrightarrow{i_W} & W^{**} \end{array}$$

(b) Show that there do *not* exist isomorphisms $i_V: V \rightarrow V^*$ such that the following diagram always commutes.

$$\begin{array}{ccc} V & \xrightarrow{i_V} & V^* \\ f \downarrow & & \uparrow f^* \\ W & \xrightarrow{i_W} & W^* \end{array}$$

Hint: There does not even exist an isomorphism $i: \mathbb{R} \rightarrow \mathbb{R}^*$ which makes the diagram commute for all linear $f: \mathbb{R} \rightarrow \mathbb{R}$.

7. A *covariant functor* from (finite dimensional) vector spaces to vector spaces is a function F which assigns to every vector space V a vector space $F(V)$ and to every linear transformation $f: V \rightarrow W$ a linear transformation $F(f): F(V) \rightarrow F(W)$, such that $F(1_V) = 1_{F(V)}$ and $F(g \circ f) = F(g) \circ F(f)$.

- (a) The “identity functor”, $F(V) = V$, $F(f) = f$ is a functor.
- (b) The “double dual functor”, $F(V) = V^{**}$, $F(f) = f^{**}$ is a functor.
- (c) The “ \mathcal{T}_k functor”, $F(V) = \mathcal{T}_k(V) = \mathcal{T}^k(V^*)$,

$$F(f)(T)(\lambda_1, \dots, \lambda_k) = T(\lambda_1 \circ f, \dots, \lambda_k \circ f)$$

is a functor.

(d) If F is any functor and $f: V \rightarrow W$ is an isomorphism, then $F(f)$ is an isomorphism.

A *contravariant functor* is defined similarly, except that $F(f): F(W) \rightarrow F(V)$ and $F(g \circ f) = F(f) \circ F(g)$. Functors of more than one argument, covariant in some and contravariant in others, may also be defined.

- (e) The “dual functor”, $F(V) = V^*$, $F(f) = f^*$ is a contravariant functor.
- (f) The “ \mathcal{T}^k functor”, $F(V) = \mathcal{T}^k(V)$, $F(f) = f^*$ is a contravariant functor.

8. (a) Let $\text{Hom}(V, W)$ denote all linear transformations from V to W . Choosing a basis for V and W , we can identify $\text{Hom}(V, W)$ with the $m \times n$ matrices, and consequently give it the metric of \mathbb{R}^{nm} . Show that a different choice of bases leads to a homeomorphic metric on $\text{Hom}(V, W)$.

(b) A functor F gives a map from $\text{Hom}(V, W)$ to $\text{Hom}(F(V), F(W))$. Call F *continuous* if this map is always continuous [using the metric in part (a)]. Show that if $\xi = \pi: E \rightarrow B$ is any vector bundle, and F is continuous, then there is a bundle $F(\xi) = \pi': E' \rightarrow B$ for which $\pi'^{-1}(p) = F(\pi^{-1}(p))$, and such that to every trivialization

$$t: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^n$$

corresponds a trivialization

$$t': \pi'^{-1}(U) \rightarrow U \times F(\mathbb{R}^n).$$

(c) The functor $\mathcal{T}_1(V) = \mathcal{T}^1(V^*) = V^{**}$ is continuous. (The bundle $\mathcal{T}_1(TM)$ is just a case of the construction in (b).)

(d) Define a continuous contravariant functor F , and show how to construct a bundle $F(\xi)$.

(e) The functor $F(V) = V^*$ is continuous. (The bundle T^*M is a special case of the construction in (d).)

Generally, the same construction can be used when F is a functor of several arguments. The bundles $\mathcal{T}_i^k(M)$ are all special cases. See the next two problems for other examples, as well as an example of a functor which is not continuous.

9. (a) Let F be a functor from \mathbf{V}^n , the class of n -dimensional vector spaces, to \mathbf{V}^k . Given $A \in \text{GL}(n, \mathbb{R})$ we can consider it as a map $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$. Then $F(A): F(\mathbb{R}^n) \rightarrow F(\mathbb{R}^n)$. Choose, once and for all, an isomorphism $F(\mathbb{R}^n) \rightarrow \mathbb{R}^k$. Then $F(A)$ can be considered as a map $h(A): \mathbb{R}^k \rightarrow \mathbb{R}^k$. Show that $h: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(k, \mathbb{R})$ is a homomorphism.

(b) How does the homomorphism h depend on the initial choice of the isomorphism $F(\mathbb{R}^n) \rightarrow \mathbb{R}^k$?

(c) Let $\mathbf{v} = (v_1, \dots, v_n)$ and $\mathbf{w} = (w_1, \dots, w_n)$ be ordered bases of V and let $\mathbf{e} = (e_1, \dots, e_n)$ be the standard basis of \mathbb{R}^n . If $\mathbf{e} \rightarrow \mathbf{v}$ denotes the isomorphism taking e_i to v_i , show that the following diagram commutes

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{\mathbf{e} \rightarrow \mathbf{v}} & V \\ A \uparrow & \nearrow \mathbf{e} \rightarrow \mathbf{w} & \\ \mathbb{R}^n & & \end{array} ,$$

where $A = (a_{ij})$ is defined by

$$w_i = \sum_{j=1}^n a_{ji} v_j.$$

After identifying $F(\mathbb{R}^n)$ with \mathbb{R}^k , this means that

$$\begin{array}{ccc} \mathbb{R}^k & \xrightarrow{F(\mathbf{e} \rightarrow \mathbf{v})} & F(V) \\ h(A) \uparrow & \nearrow F(\mathbf{e} \rightarrow \mathbf{w}) & \\ \mathbb{R}^k & & \end{array}$$

also commutes. This suggests a way of proving the following.

THEOREM. If $h: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(k, \mathbb{R})$ is any homomorphism, there is a functor $F_h: \mathbf{V}^n \rightarrow \mathbf{V}^k$ such that the homomorphism defined in part (a) is equal to h .

(d) For $q, q' \in \mathbb{R}^k$, define

$$(\mathbf{v}, q) \sim (\mathbf{w}, q')$$

if $q = h(A)q'$ where $w_i = \sum_{j=1}^n a_{ji} v_j$. Show that \sim is an equivalence relation, and that every equivalence class contains exactly one element (\mathbf{v}, q) for a given \mathbf{v} . We will denote the equivalence class of (\mathbf{v}, q) by $[\mathbf{v}, q]$.

(e) Show that the operations

$$\begin{aligned} [\mathbf{v}, q_1] + [\mathbf{v}, q_2] &= [\mathbf{v}, q_1 + q_2] \\ a \cdot [\mathbf{v}, q] &= [\mathbf{v}, aq] \end{aligned}$$

are well-defined operations making the set of all equivalence classes into a k -dimensional vector space $F_h(V)$.

(f) If $V, W \in \mathbf{V}^n$ and $f: V \rightarrow W$, choose ordered bases \mathbf{v}, \mathbf{w} , define A by $f(v_i) = \sum_{j=1}^n a_{ji} w_j$, and define

$$F_h(f)[\mathbf{v}, q] = [\mathbf{w}, h(A)(q)].$$

Show that this is a well-defined linear transformation, that F_h is a functor, and that $F_h(A) = h(A)$ when we identify $F_h(\mathbb{R}^n)$ with \mathbb{R}^k by $[\mathbf{e}, q] \mapsto q$.

(g) Let $\alpha: \mathbb{R} \rightarrow \mathbb{R}$ be a non-continuous homomorphism (compare page 380), and let $h: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(1, \mathbb{R}) = \mathbb{R}$ be $h(A) = \alpha(\det A)$. Then $F_h: \mathbf{V}^n \rightarrow \mathbf{V}^1$ is a non-continuous functor.

10. In classical tensor analysis there are, in addition to mixed tensor fields, other “quantities” which are defined as sets of functions which transform according to yet other rules. These new rules are of the form

$$A' = A \text{ operated on by } h \left(\frac{\partial x^\alpha}{\partial x'^\beta} \right).$$

For example, assignments of a single function a to each coordinate system x such that the function a' assigned to x' satisfies

$$a' = \det \left(\frac{\partial x^i}{\partial x'^j} \right) \cdot a$$

are called (even) scalar densities; assignments for which

$$a' = \left| \det \left(\frac{\partial x^i}{\partial x'^j} \right) \right| \cdot a$$

are called odd scalar densities. The Theorem in Problem 9 allows us to construct a bundle whose sections correspond to these classical entities (later we will have a more illuminating way):

(a) Let $h: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(1, \mathbb{R})$ take A into multiplication by $\det A$. Let F_h be the functor given by the Theorem, and consider the 1-dimensional bundle $F_h(TM)$ obtained by replacing each fibre M_p with $F_h(M_p)$. If (x, U) is a coordinate system, then

$$\alpha_x(p) = \left[\left(\frac{\partial}{\partial x^1} \Big|_p, \dots, \frac{\partial}{\partial x^n} \Big|_p \right), 1 \right] \in F_h(M_p)$$

is non-zero, so every section on U can be expressed as $a \cdot \alpha_x$ for a unique function a . If x' is another coordinate system and $a \cdot \alpha_x = a' \cdot \alpha_{x'}$, show that

$$a' = \det \left(\frac{\partial x^i}{\partial x'^j} \right) \cdot a.$$

(b) If, instead, h takes A into multiplication by $|\det A|$, show that the corresponding equation is

$$a' = \left| \det \left(\frac{\partial x^i}{\partial x'^j} \right) \right| \cdot a.$$

(c) For this h , show that a non-zero element of $F_h(V)$ determines an orientation for V . Conclude that the bundle of odd scalar densities is *not* trivial if M is not orientable.

(d) We can identify $\mathcal{T}_l^k(\mathbb{R}^n)$ with $\mathbb{R}^{n^{k+l}}$ by taking

$$e^*_{i_1} \otimes \cdots \otimes e^*_{i_k} \otimes e_{j_1} \otimes \cdots \otimes e_{j_l} \mapsto (i_1, \dots, i_k, j_1, \dots, j_l)^{\text{th}} \text{ basis vector of } \mathbb{R}^{n^{k+l}}.$$

Recall that if $f: V \rightarrow V$, we define $\mathcal{T}_l^k(f): \mathcal{T}_l^k(V) \rightarrow \mathcal{T}_l^k(V)$ by

$$\mathcal{T}_l^k(f)(T)(v_1, \dots, v_k, \lambda_1, \dots, \lambda_l) = T(f(v_1), \dots, f(v_k), \lambda_1 \circ f, \dots, \lambda_l \circ f).$$

Given $A \in \text{GL}(n, \mathbb{R})$, we can consider it as a map $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$. Then $\mathcal{T}_l^k(A): \mathcal{T}_l^k(\mathbb{R}^n) \rightarrow \mathcal{T}_l^k(\mathbb{R}^n)$ determines an element $\mathcal{T}_l^k(A)$ of $\text{GL}(n^{k+l}, \mathbb{R})$. Let $h: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(n^{k+l}, \mathbb{R})$ be defined by

$$h(A) = (\det A)^w \mathcal{T}_l^k(A) \quad w \text{ an integer.}$$

The bundle $F(TM)$ is called the bundle of (even) relative tensors of type $\binom{k}{l}$ and weight w . For $k = l = 0$ we obtain the bundle of (even) relative scalars of weight w [the (even) scalar densities are the (even) relative scalars of weight 1]. If $(\det A)^w$ is replaced by $|\det A|^w$ (w any real number), we obtain the bundle of odd relative tensors of type $\binom{k}{l}$ and weight w . Show that the transformation law for the components of sections of these bundles is

$$A'^{\beta_1 \dots \beta_l}_{\alpha_1 \dots \alpha_k} = \left[\det \left(\frac{\partial x^i}{\partial x'^j} \right) \right]^w \sum_{\substack{i_1, \dots, i_k \\ j_1, \dots, j_l}} A^{j_1 \dots j_l}_{i_1 \dots i_k} \frac{\partial x^{i_1}}{\partial x'^{\alpha_1}} \cdots \frac{\partial x^{i_k}}{\partial x'^{\alpha_k}} \frac{\partial x'^{\beta_1}}{\partial x^{j_1}} \cdots \frac{\partial x'^{\beta_l}}{\partial x^{j_l}}$$

(or the same formula with $\det(\partial x^i / \partial x'^j)$ replaced by $|\det(\partial x^i / \partial x'^j)|$).

(e) Define

$$\varepsilon_{i_1 \dots i_n} = \begin{cases} +1 & \text{if } i_1, \dots, i_n \text{ is an even permutation of } 1, \dots, n \\ -1 & \text{if } i_1, \dots, i_n \text{ is an odd permutation of } 1, \dots, n \\ 0 & \text{if } i_\alpha = i_\beta \text{ for some } \alpha \neq \beta. \end{cases}$$

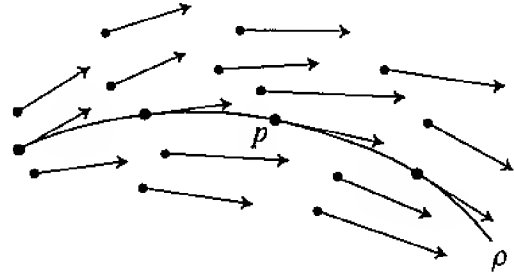
Show that there is a covariant relative tensor of weight -1 with these components in every coordinate system. Also show that $\varepsilon^{i_1 \dots i_n} = \varepsilon_{i_1 \dots i_n}$ are the components in every coordinate system of a certain contravariant relative tensor of weight 1. (See Problem 7-12 for a geometric interpretation of these relative tensors.)

CHAPTER 5

VECTOR FIELDS AND DIFFERENTIAL EQUATIONS

We return to a more detailed study of the tangent bundle TM , and its sections, i.e., vector fields. Let X be a vector field defined in a neighborhood of $p \in M$. We would like to know if there is a curve $\rho: (-\varepsilon, \varepsilon) \rightarrow M$ through p whose tangent vectors coincide with X , that is, a curve ρ with

$$\begin{aligned}\rho(0) &= p \\ \rho_* \left(\frac{d}{dt} \Big|_t \right) &= \frac{d\rho}{dt} \Big|_t = X(\rho(t)).\end{aligned}$$



Since this is a local question, we wish to introduce a coordinate system (x, U) around p and transfer the vector field X to $x(U) \subset \mathbb{R}^n$. Recall that, in general, $\alpha_* X$ does not make sense for C^∞ functions $\alpha: M \rightarrow N$. However, if α is a diffeomorphism, then we define

$$(\alpha_* X)_q = \alpha_* (X_{\alpha^{-1}(q)}) \quad [\text{i.e., } = \alpha_{*\alpha^{-1}(q)}(X_{\alpha^{-1}(q)})].$$

It is not hard to check (Problem 1) that $\alpha_* X$ is C^∞ on $\alpha(M)$. In particular, we have a vector field $x_* X$ on $x(U) \subset \mathbb{R}^n$. There is a function $f: x(U) \rightarrow \mathbb{R}^n$ with

$$(x_* X)_q = f(q)_q \in \mathbb{R}^n_q,$$

i.e., $(x_* X)_q$ has “components” $f^1(q), \dots, f^n(q)$. Consider the curve $c = x \circ \rho$. The condition

$$\frac{d\rho}{dt} = X(\rho(t))$$

means that

$$\rho_* \left(\frac{d}{dt} \Big|_t \right) = X(\rho(t));$$

hence

$$\begin{aligned}\left.\frac{dc}{dt}\right|_t &= x_*\rho_*\left(\left.\frac{d}{dt}\right|_t\right) = x_*(X(\rho(t))) = (x_*X)_{x(\rho(t))} \\ &= (x_*X)_{c(t)}.\end{aligned}$$

If we use $c'(t)$ to denote the ordinary derivative of the \mathbb{R}^n -valued function c , then this equation finally becomes simply

$$c'(t) = f(c(t)).$$

This is a simple example of a differential equation for a function $c: \mathbb{R} \rightarrow \mathbb{R}^n$, which may also be considered as a system of n differential equations for the functions c^i ,

$$c^{i'}(t) = f^i(c^1(t), \dots, c^n(t)) \quad i = 1, \dots, n.$$

We also want the “initial conditions”

$$c^i(0) = x^i(p).$$

Solving a differential equation used to be described as “integrating” the equation (the process is integration when the equation has the special form $c'(t) = f(t)$ for $f: \mathbb{R} \rightarrow \mathbb{R}$, a form to which our particular equations never reduce); solutions were consequently called “integrals” of the equation. Part of this terminology is still preserved. A curve $\rho: (-\varepsilon, \varepsilon) \rightarrow M$ with

$$\begin{aligned}\rho(0) &= p \\ \frac{d\rho}{dt} &= X(\rho(t))\end{aligned}$$

is called an **integral curve** for X with **initial condition** $\rho(0) = p$. Similar terminology is applied, of course, to the differential equations one obtains upon introducing a coordinate system. For quite some time, we will work entirely in Euclidean space, and for a while x, y , etc., will denote points of \mathbb{R}^n . If $U \subset \mathbb{R}^n$ is open and $f: U \rightarrow \mathbb{R}^n$, then a curve $c: (-\varepsilon, \varepsilon) \rightarrow M$ with

$$\begin{aligned}c(0) &= x & x &\in U \\ c'(t) &= f(c(t))\end{aligned}$$

is called an **integral curve** for f with **initial condition** $c(0) = x$.

Before stating the main theorem about the existence and uniqueness of such integral curves, we consider some special cases.

The equation for a curve c with range \mathbb{R} ,

$$c'(t) = -[c(t)]^2,$$

which would be written classically in terms of a function $y: \mathbb{R} \rightarrow \mathbb{R}$ as

$$\frac{dy}{dx} = -y^2,$$

is the special case $f(a) = -a^2$. The standard method of solving this equation is to write

$$\begin{aligned}\frac{dy}{-y^2} &= dx \\ \int \frac{dy}{-y^2} &= \int dx \\ \frac{1}{y} &= x + C \\ y &= \frac{1}{x + C}.\end{aligned}$$

Thus the curves

$$c(t) = \frac{1}{t + C}$$

are supposed to be solutions. This can be checked directly if you don't believe the above manipulations. (They really do make sense; the equation in question asserts that $y' = f \circ y$, so

$$\left(\frac{1}{f} \circ y\right) \cdot y' = 1;$$

hence, if $F' = 1/f$, then

$$\begin{aligned}(F \circ y)' &= 1 \\ F(y(x)) &= x + C\end{aligned}$$

for some C .) To obtain the initial conditions $c(0) = a$, we must take

$$c(t) = \frac{1}{t + 1/a}.$$

This works in all cases except $a = 0$. In this case, the correct solution is

$$c(t) = 0 \quad \text{for all } t$$

(which we missed by dividing by y). In terms of vector fields, the curves c are the integral curves of

$$X(a) = -a^2 \frac{d}{dt}.$$



Notice that no integral curve, except $c(t) = 0$, can be defined for all t , even though X is defined on all of \mathbb{R} . It might be thought that this somehow reflects the fact that $X(0) = 0$, but this has nothing to do with the case. For $a > 0$, the curve $c(t) = 1/(t + 1/a)$ is defined for all large t , and as $t \rightarrow \infty$ it approaches, but never reaches, 0. On the other hand, as $t \rightarrow -1/a$ the curve escapes to infinity because the vector field gets big too fast. This will continue to be true even if we modify the vector field near 0 so that it is never 0.

Another phenomenon is illustrated by the equation

$$c'(t) = c(t)^{2/3},$$

written classically as

$$\frac{dy}{dx} = y^{2/3}.$$

There are two different solutions with the initial condition $c(0) = 0$, namely

- (1) $c(t) = 0$ for all t ,
- (2) $c(t) = \frac{1}{27}t^3$ for all t .

In this case, the function f , given by $f(a) = a^{2/3}$, is *not* differentiable. Uniqueness will always be insured when $f: U \rightarrow \mathbb{R}^n$ is C^1 , but it can also be obtained with a rather less stringent condition. We say that the function f satisfies a **Lipschitz condition** on U if there is some K such that

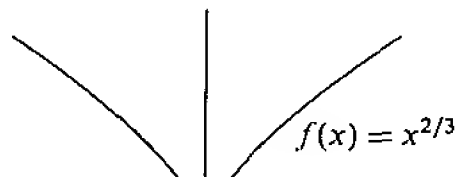
$$|f(x) - f(y)| \leq K|x - y| \quad \text{for all } x, y \in U.$$

Notice that $f(a) = a^{2/3}$ is not Lipschitz; in fact, there is no K with

$$|f(x) - f(0)| \leq K|x|$$

for x near 0, since

$$\frac{x^{2/3}}{x} = x^{-1/3} \rightarrow \pm\infty \quad \text{as } x \rightarrow 0^\pm.$$



A Lipschitz function is clearly continuous, but not necessarily differentiable (for example, $f(x) = |x|$). On the other hand, a C^1 function is locally Lipschitz, that is, it satisfies a Lipschitz condition in a neighborhood of each point—this follows from Lemma 2-5. A Lipschitz function is also clearly bounded on any bounded set.

The basic existence and uniqueness theorem for differential equations depends on a simple lemma about complete metric spaces.

1. THEOREM (THE CONTRACTION LEMMA). Let (M, ρ) be a non-empty complete metric space, and let $f: M \rightarrow M$ be a “contraction”, that is, suppose there is some $C < 1$ such that

$$\rho(f(x), f(y)) \leq C\rho(x, y) \quad \text{for all } x, y \in M.$$

Then there is a unique $x \in M$ such that $f(x) = x$ (the function f has a unique “fixed point”).

PROOF. Notice that f is clearly continuous. Let $x_0 \in M$ and define a sequence $\{x_n\}$ inductively by

$$x_{n+1} = f(x_n),$$

i.e.,

$$x_{n+1} = f^n(x_0) = \underbrace{f \circ f \circ \cdots \circ f}_{n \text{ times}}(x_0).$$

Then an easy induction argument shows that

$$\rho(x_n, x_{n+1}) \leq C^n \rho(x_0, x_1).$$

Thus

$$\begin{aligned} \rho(x_n, x_{n+k}) &\leq \rho(x_n, x_{n+1}) + \cdots + \rho(x_{n+k-1}, x_{n+k}) \\ &\leq (C^n + \cdots + C^{n+k-1})\rho(x_0, x_1). \end{aligned}$$

Since $C < 1$, the sum $\sum_{n=0}^{\infty} C^n$ converges, so $C^n + \cdots + C^{n+k-1} \rightarrow 0$ as $n \rightarrow \infty$. Thus the sequence $\{x_n\}$ is Cauchy, so there is some x with

$$x = \lim_{n \rightarrow \infty} x_n.$$

Continuity of f then shows that

$$f(x) = \lim_{n \rightarrow \infty} f(x_n) = \lim_{n \rightarrow \infty} x_{n+1} = x. \quad \spadesuit$$

We are going to apply the Contraction Lemma to certain spaces of functions. Recall that if (M, ρ) is a metric space and X is compact, then the set of all continuous functions $f: X \rightarrow M$ is a metric space if we define the metric σ by

$$\sigma(f, g) = \sup_{x \in X} \rho(f(x), g(x)).$$

If M is bounded, then we do not even need X to be compact. Moreover, if M is complete, then the new metric space is also complete; this is basically just the theorem that the uniform limit of continuous functions is continuous, plus the fact that each $\lim_{x \rightarrow \infty} f_n(x)$ exists since M is complete. In particular, if M is a compact subset of \mathbb{R}^n , then the set of all continuous functions $f: X \rightarrow M$ is complete with the metric

$$\sigma(f, g) = \|f - g\|, \quad \text{where } \|f\| = \sup_{x \in X} |f(x)|.$$

Our basic strategy in solving differential equations will be to replace differentiable functions and derivatives by continuous functions and integrals. If $U \subset \mathbb{R}^n$ and $f: U \rightarrow \mathbb{R}^n$ is continuous, then a continuous function $\alpha: (-b, b) \rightarrow U$, defined on some interval around 0, clearly satisfies

$$(1) \quad \begin{aligned} \alpha'(t) &= f(\alpha(t)) \\ \alpha(0) &= x \end{aligned}$$

if it satisfies the integral equation

$$(2) \quad \alpha(t) = x + \int_0^t f(\alpha(u)) du,$$

where the integral of an \mathbb{R}^n -valued function is defined by integrating each component function separately. Conversely, if α satisfies (1), then α is differentiable, hence continuous; thus $\alpha' = f \circ \alpha$ is continuous, so

$$\alpha(t) - x = \alpha(t) - \alpha(0) = \int_0^t \alpha'(u) du = \int_0^t f(\alpha(u)) du.$$

For the proof of the basic theorem, we need only one simple estimate. If a continuous function $f: [a, b] \rightarrow \mathbb{R}^n$ satisfies $|f| \leq K$, then

$$\left| \int_a^b f(u) du \right| \leq K(b - a).$$

To prove this, we note that it is true for constant functions, hence for step functions, and thus for continuous functions, which are uniform limits on $[a, b]$ of step functions.

2. THEOREM. Let $f: U \rightarrow \mathbb{R}^n$ be any function, where $U \subset \mathbb{R}^n$ is open. Let $x_0 \in U$ and let $a > 0$ be a number such that the closed ball $\bar{B}_{2a}(x_0)$, of radius $2a$ and center x_0 , is contained in U . Suppose that

$$(1) |f| \leq L \text{ on } \bar{B}_{2a}(x_0)$$

$$(2) |f(x) - f(y)| \leq K|x - y| \text{ for } x, y \in \bar{B}_{2a}(x_0).$$

Choose $b > 0$ so that

$$(3) b \leq a/L$$

$$(4) b < 1/K.$$

Then for each $x \in \bar{B}_a(x_0)$ there is a unique $\alpha_x: (-b, b) \rightarrow U$ such that

$$\alpha_x'(t) = f(\alpha_x(t))$$

$$\alpha_x(0) = x.$$

PROOF. Choose $x \in \bar{B}_a(x_0)$, which will be fixed for the remainder of the proof. Let

$$M = \{\text{continuous } \alpha: (-b, b) \rightarrow \bar{B}_{2a}(x_0)\}.$$

Then M is a complete metric space. For each $\alpha \in M$, define a curve $S\alpha$ on $(-b, b)$ by

$$S\alpha(t) = x + \int_0^t f(\alpha(u)) du$$

(the integral exists since f is continuous on $\bar{B}_{2a}(x_0)$). The curve $S\alpha$ is clearly continuous. Moreover, for any $t \in (-b, b)$ we have

$$\begin{aligned} |S\alpha(t) - x| &= \left| \int_0^t f(\alpha(u)) du \right| \\ &< bL \quad \text{by (1)} \\ &\leq a \quad \text{by (3).} \end{aligned}$$

Since $|x - x_0| \leq a$, it follows that $|S\alpha(t) - x_0| < 2a$, for all $t \in (-b, b)$, so

$$(*) \quad S\alpha(t) \in B_{2a}(x_0) \subset \bar{B}_{2a}(x_0) \quad \text{for } t \in (-b, b).$$

Thus $S: M \rightarrow M$.

Now suppose $\alpha, \beta \in M$. Then

$$\begin{aligned} \|S\alpha - S\beta\| &= \sup_t \left| \int_0^t f(\alpha(u)) - f(\beta(u)) du \right| \\ &< bK \sup_{-b < u < b} |\alpha(u) - \beta(u)| \quad \text{by (2)} \\ &= bK\|\alpha - \beta\|. \end{aligned}$$

Since we chose $bK < 1$ (by (4)), this shows that $S: M \rightarrow M$ is a contraction. Hence S has a unique fixed point:

There is a unique $\alpha: (-b, b) \rightarrow \bar{B}_{2a}(x_0)$ with

$$\alpha(t) = x + \int_0^t f(\alpha(u)) du.$$

This, alas, is not quite what the theorem states. Having used the elegant Contraction Lemma, we pay for it by finishing off with a finicky detail:

The map α is the unique $\beta: (-b, b) \rightarrow U$ satisfying

$$\beta(t) = x + \int_0^t f(\beta(u)) du.$$

Reason: We claim that any such β actually lies in $\bar{B}_{2a}(x_0)$, in fact, in $B_{2a}(x_0)$. Consider first numbers $t > 0$. We have already seen (statement (*)) that for each t with $0 \leq t < b$,

$$(**) \quad \beta(t) = x + \int_0^t f(\beta(u)) du \quad \text{is in } B_{2a}(x_0) \quad [\text{the open ball}]$$

provided that

$$\beta(u) \in \bar{B}_{2a}(x_0) \quad \text{for all } u \text{ with } 0 \leq u < t,$$

so certainly if

$$\beta(u) \in B_{2a}(x_0) \quad \text{for all } u \text{ with } 0 \leq u \leq t.$$

We can now use a simple least upper bound argument. Let

$$A = \{t : 0 \leq t < b \text{ and } \beta(u) \in B_{2a}(x_0) \text{ for } 0 \leq u < t\}.$$

Let $\alpha = \sup A$. Suppose $\alpha < b$. We clearly have $\beta(u) \in B_{2a}(x_0)$ for $0 \leq u < \alpha$. So $\beta(\alpha) \in B_{2a}(x_0)$, by (**). This clearly implies that $\beta(\alpha + s) \in B_{2a}(x_0)$ for sufficiently small $s > 0$, which contradicts the fact that $\alpha = \sup A$. So it must be that $\sup A = b$. A similar argument works for $-b < t \leq 0$.

To sum up, the unique fixed point α_x of the map S is the unique curve with the desired properties. ♦

Notice that solutions of the differential equation

$$\alpha'(t) = f(\alpha(t))$$

remain solutions under additive changes of parameter; that is, if

$$\beta(t) = \alpha(t_0 + t),$$

then

$$\beta'(t) = \alpha'(t_0 + t) = f(\alpha(t_0 + t)) = f(\beta(t)).$$

This remark allows us to extend the uniqueness part of Theorem 2.

3. THEOREM. Suppose $f: U \rightarrow \mathbb{R}^n$ is locally Lipschitz, that is, around each point there is a ball on which f satisfies condition (2) of Theorem 2 for some K (and hence also condition (1) for some L). Let $x \in U$ and let α_1, α_2 be two maps on some open interval I with $\alpha_1(I), \alpha_2(I) \subset U$ and

$$\begin{aligned} \alpha_i'(t) &= f(\alpha_i(t)) \\ \alpha_i(0) &= x \end{aligned} \quad i = 1, 2.$$

Then $\alpha_1 = \alpha_2$ on I .

PROOF. Suppose $\alpha_1(t_0) = \alpha_2(t_0)$ for some $t_0 \in I$. If we define

$$\beta_i(t) = \alpha_i(t_0 + t),$$

then the functions β_i satisfy the same differential equation, $\beta_i'(t) = f(\beta_i(t))$, and have the same initial condition $\beta_i(0) = \alpha_1(t_0) = \alpha_2(t_0) \in U$. Hence $\beta_1(t) = \beta_2(t)$ for sufficiently small t , by Theorem 1. Thus the set

$$\{t \in I : \alpha_1(t) = \alpha_2(t)\}$$

is open. It is clearly also closed and non-empty, so it equals I . ♦

We now revert to the situation in Theorem 2. We will write $\alpha_x(t)$ as $\alpha(t, x)$, so that we have a map

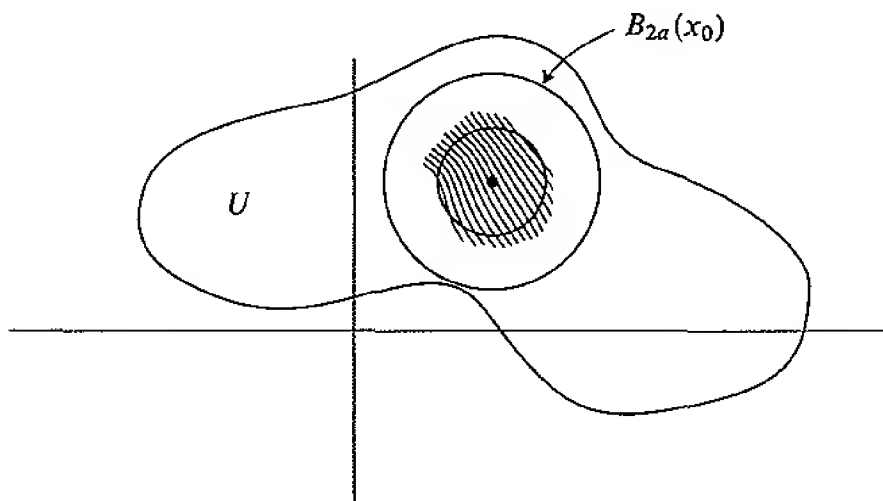
$$\alpha: (-b, b) \times B_a(x_0) \rightarrow U$$

satisfying

$$\begin{aligned} \alpha(0, x) &= x \\ \frac{d}{dt}\alpha(t, x) &= f(\alpha(t, x)) \end{aligned}$$

[i.e., $D_1\alpha(t, x) = f(\alpha(t, x))$], but we will frequently use $\partial/\partial t$ or d/dt in this

discussion]. This map α is called a local flow for f in $(-b, b) \times B_a(x_0)$. To picture this map α , the best we can do is to draw the images of the integral



curves α_x . If $y = \alpha_x(t_0)$, then the integral curve α_x with the initial condition $\alpha_x(0) = x$ differs from the integral curve α_y with initial condition $\alpha_y(0) = y$ only by a change of parameter, so the two images overlap. For each fixed x , the map $t \mapsto \alpha(t, x)$ for $-b < t < b$ lies along *part* of the curve through x . On the other hand, if we fix t , then the map

$$x \mapsto \alpha(t, x)$$

gives the result of pushing each x along the integral curve through it, for a time interval of t . To focus attention on this map, we denote it by ϕ_t :

$$\phi_t(x) = \alpha(t, x) \quad [= \alpha_x(t)].$$

This map ϕ_t is always continuous. In fact, the whole flow α is continuous (as a function of both t and x):

4. THEOREM. If $f: U \rightarrow \mathbb{R}^n$ is locally Lipschitz, then the flow

$$\alpha: (-b, b) \times B_a(x_0) \rightarrow U$$

given by Theorem 2 is continuous.

PROOF. Let us denote the map S defined in the proof of Theorem 2 by S_x , to indicate explicitly the role of x . Then

$$\|\alpha_x - S_y \alpha_x\| = \|S_x \alpha_x - S_y \alpha_x\| = |x - y|.$$

Recall that

$$\|S\alpha - S\beta\| \leq bK\|\alpha - \beta\|.$$

If S_y^n denotes the n -fold iterate of S_y , then

$$\begin{aligned}\|\alpha_x - S_y^n \alpha_x\| &\leq \|\alpha_x - S_y \alpha_x\| + \|S_y \alpha_x - S_y^2 \alpha_x\| + \cdots + \|S_y^{n-1} \alpha_x - S_y^n \alpha_x\| \\ &\leq (1 + bK + \cdots + (bK)^{n-1})|x - y| \leq \frac{1}{1 - bK}|x - y|.\end{aligned}$$

Recall also that in Theorem 1 the fixed point α_y of S_y is the limit of $S_y^n \alpha$ for any α . Hence $\alpha_y = \lim_{n \rightarrow \infty} S_y^n \alpha_x$, so we obtain

$$\|\alpha_x - \alpha_y\| \leq \frac{1}{1 - bK}|x - y|.$$

Since $\|\alpha_x - \alpha_y\| = \sup_t |\alpha(t, x) - \alpha(t, y)|$, this certainly proves continuity of α . ♦

If additional conditions are placed upon the map f , then further smoothness conditions can be proved for α . In fact,

If $f: U \rightarrow \mathbb{R}^n$ is C^k , then the flow $\alpha: (-b, b) \times B_a(x_0) \rightarrow U$ is also C^k .

Unfortunately, this is a very hard theorem. A clean exposition of the classical proof is given in Lang's *Introduction to Differentiable Manifolds* (2nd ed.), and a recently discovered proof can be found in Lang, *Real and Functional Analysis* (3rd ed.), pp. 371–379. In order to read this high-powered proof, you must first learn the elements of Banach spaces, including the Hahn-Banach theorem, and then read about differential calculus in Banach spaces, including the inverse and implicit function theorems (*Real and Functional Analysis*, pp. 360–365), but this is probably easier than reading the classical proof (and, besides, when you're finished you'll also know about Banach spaces, and differential calculus in Banach spaces).

We will just accept this fact. Notice that the maps ϕ_t are consequently C^∞ if f is C^∞ .

Since the map

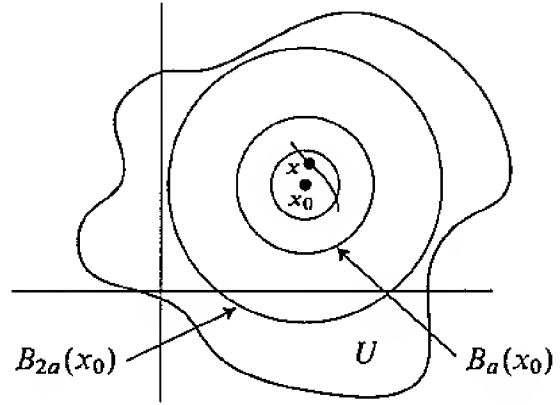
$$\alpha: (-b, b) \times B_a(x_0) \rightarrow U$$

satisfies $\alpha(0, x) = x$, we have

$$\alpha: \{0\} \times \bar{B}_{a/2}(x_0) \rightarrow \bar{B}_{a/2}(x_0) \subset B_a(x_0).$$

Continuity of α and compactness of $\{0\} \times \bar{B}_{a/2}(x_0)$ imply that there is some $\varepsilon > 0$ such that

$$\alpha: (-\varepsilon, \varepsilon) \times B_{a/2}(x_0) \rightarrow B_a(x_0).$$



[If $x \in B_{a/2}(x_0)$, then the integral curve with initial condition x stays in $B_a(x_0)$ for $|t| < \varepsilon$.]

So if $|s| < \varepsilon$, and $x \in B_{a/2}(x_0)$, then the point $\alpha(s, x) \in B_a(x_0)$, so we can also define

$$\gamma(t) = \alpha(t, \alpha(s, x)) \quad |t| < \varepsilon.$$

This satisfies

$$\begin{aligned} \gamma'(t) &= f(\gamma(t)) \\ \gamma(0) &= \alpha(s, x). \end{aligned}$$

We have also noted that

$$\beta(t) = \alpha(s + t, x), \quad \text{defined for } |s + t| < \varepsilon,$$

satisfies

$$\begin{aligned} \beta'(t) &= f(\beta(t)) \\ \beta(0) &= \alpha(s, x). \end{aligned}$$

Consequently, $\beta(t) = \alpha(t, \alpha(s, x))$ for $|t| < \varepsilon$. In other words,

$$\text{if } |s|, |t|, |s + t| < \varepsilon, \quad \text{then } \alpha(t, \alpha(s, x)) = \alpha(s + t, x).$$

If we now let $\phi_t: B_{a/2}(x_0) \rightarrow \mathbb{R}^n$ be $\phi_t(x) = \alpha(t, x)$ for $x \in B_{a/2}(x_0)$, we can say:

$$\begin{aligned} \text{if } |s|, |t|, |s + t| < \varepsilon \text{ and } x, \phi_t(x) \in B_{a/2}(x_0), \text{ then} \\ \phi_s(\phi_t(x)) &= \phi_{s+t}(x). \end{aligned}$$

Roughly speaking, $\phi_{t+s} = \phi_t \circ \phi_s = \phi_s \circ \phi_t$. This shows, in particular, that for $|s| < \varepsilon$ each ϕ_s is a diffeomorphism, with inverse $\phi_s^{-1} = \phi_{-s}$. Everything we have said, since it is local, can be resaid, without requiring any more proof, on a manifold.

5. THEOREM. Let X be a C^∞ vector field on M , and let $p \in M$. Then there is an open set V containing p and an $\varepsilon > 0$, such that there is a unique collection of diffeomorphisms $\phi_t: V \rightarrow \phi_t(V) \subset M$ for $|t| < \varepsilon$ with the following properties:

(1) $\phi: (-\varepsilon, \varepsilon) \times V \rightarrow M$, defined by $\phi(t, p) = \phi_t(p)$, is C^∞ .

(2) If $|s|, |t|, |s+t| < \varepsilon$, and $q, \phi_t(q) \in V$, then

$$\phi_{s+t}(q) = \phi_s \circ \phi_t(q).$$

(3) If $q \in V$, then X_q is the tangent vector at $t = 0$ of the curve $t \mapsto \phi_t(q)$.

The examples given previously show that we cannot expect ϕ_t to be defined for all t , or on all of M . In one case however, this can be attained. The support of a vector field X is just the closure of $\{p \in M : X_p \neq 0\}$.

6. THEOREM. If X has compact support (in particular, if M is compact), then there are diffeomorphisms $\phi_t: M \rightarrow M$ for all $t \in \mathbb{R}$ with properties (1), (2), (3).

PROOF. Cover support X by a finite number of open sets V_1, \dots, V_n given by Theorem 5 with corresponding $\varepsilon_1, \dots, \varepsilon_n$ and diffeomorphisms ϕ_t^i . Let $\varepsilon = \min(\varepsilon_1, \dots, \varepsilon_n)$. Notice that by uniqueness, $\phi_t^i(q) = \phi_t^j(q)$ for $q \in V_i \cap V_j$. So we can define

$$\phi_t(q) = \begin{cases} \phi_t^i(q) & \text{if } q \in V_i \\ q & \text{if } q \notin \text{support } X. \end{cases}$$

Clearly $\phi: (-\varepsilon, \varepsilon) \times M \rightarrow M$ is C^∞ , and $\phi_{t+s} = \phi_t \circ \phi_s$ if $|t|, |s|, |t+s| < \varepsilon$, and each ϕ_t is a diffeomorphism.

To define ϕ_t for $|t| \geq \varepsilon$, write

$$t = k(\varepsilon/2) + r \quad \text{with } k \text{ an integer, and } |r| < \varepsilon/2.$$

Let

$$\phi_t = \begin{cases} \phi_{\varepsilon/2} \circ \dots \circ \phi_{\varepsilon/2} \circ \phi_r & [\phi_{\varepsilon/2} \text{ iterated } k \text{ times}] & \text{for } k \geq 0 \\ \phi_{-\varepsilon/2} \circ \dots \circ \phi_{-\varepsilon/2} \circ \phi_r & [\phi_{-\varepsilon/2} \text{ iterated } -k \text{ times}] & \text{for } k < 0. \end{cases}$$

It is easy to check that this is the desired $\{\phi_t\}$. ♦

The unique collection $\{\phi_t\}$ given by Theorem 6, or more precisely, the map $t \mapsto \phi_t$ from \mathbb{R} to the group of all diffeomorphisms of M , is called a 1-parameter group of diffeomorphisms, and is said to be generated by X . In the local case of Theorem 5, we obtain a “local 1-parameter group of local diffeomorphisms”. The vector field X is sometimes called the “infinitesimal generator” of $\{\phi_t\}$ (vector fields used to be called “infinitesimal transformations”).

Condition (3) in Theorem 5 can be rephrased in terms of the action of X_q on a C^∞ function $f: M \rightarrow \mathbb{R}$. Recall that

$$\frac{dc}{dt}(f) = \frac{df(c(t))}{dt} = (f \circ c)'(t).$$

Thus, to say that X_q is the tangent vector at $t = 0$ of the curve $t \mapsto \phi_t(q)$ amounts to saying that

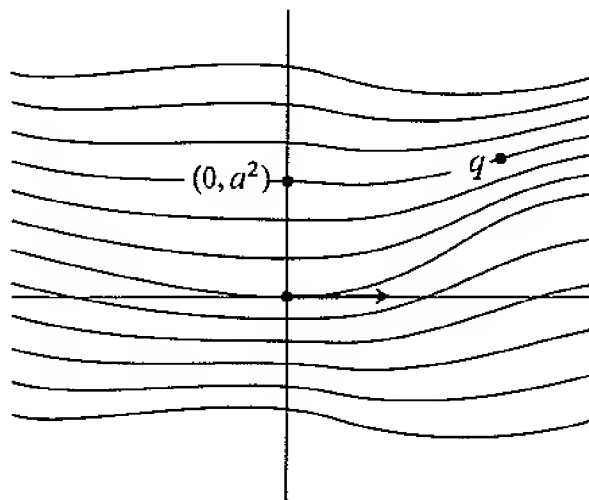
$$(Xf)(q) = X_q f = \lim_{h \rightarrow 0} \frac{f(\phi_h(q)) - f(q)}{h}.$$

This equation will be used very frequently. The first use is to derive a corollary of Theorem 5 which allows us to simplify many calculations involving vector fields, and which also has important theoretical uses.

7. THEOREM. Let X be a C^∞ vector field on M with $X(p) \neq 0$. Then there is a coordinate system (x, U) around p such that

$$X = \frac{\partial}{\partial x^1} \quad \text{on } U.$$

PROOF. It is easy to see that we can assume $M = \mathbb{R}^n$ (with the standard coordinate system t^1, \dots, t^n , say), and $p = 0 \in \mathbb{R}^n$. Moreover, we can assume that $X(0) = \partial/\partial t^1|_0$. The idea of the proof is that in a neighborhood of 0 there is a unique integral curve through each point $(0, a^2, \dots, a^n)$; if q lies on the integral



curve through this point, we will use a^2, \dots, a^n as the last $n-1$ coordinates of q and the time interval it takes the curve to get to q as the first coordinate. To do this, let X generate ϕ_t and consider the map χ defined on a neighborhood of 0 in \mathbb{R}^n by

$$\chi(a^1, \dots, a^n) = \phi_{a^1}(0, a^2, \dots, a^n).$$

We compute that for $a = (a^1, \dots, a^n)$,

$$\begin{aligned} \chi_* \left(\frac{\partial}{\partial t^1} \Big|_a \right) (f) &= \frac{\partial}{\partial t^1} \Big|_a (f \circ \chi) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(\chi(a^1 + h, a^2, \dots, a^n)) - f(\chi(a))] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(\phi_{a^1+h}(0, a^2, \dots, a^n)) - f(\chi(a))] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(\phi_h(\chi(a))) - f(\chi(a))] \\ &= (Xf)(\chi(a)). \end{aligned}$$

Moreover, for $i > 1$ we can at least compute

$$\begin{aligned} \chi_* \left(\frac{\partial}{\partial t^i} \Big|_0 \right) (f) &= \frac{\partial}{\partial t^i} \Big|_0 (f \circ \chi) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(\chi(0, \dots, h, \dots, 0)) - f(0)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(0, \dots, h, \dots, 0) - f(0)] \\ &= \frac{\partial f}{\partial t^i} \Big|_0. \end{aligned}$$

Since $X(0) = \partial/\partial t^1|_0$ by assumption, this shows that $\chi_{*0} = I$ is non-singular. Hence $x = \chi^{-1}$ may be used as a coordinate system in a neighborhood of 0. This is the desired coordinate system, for it is easy to see that the equation $\chi_*(\partial/\partial t^1) = X \circ \chi$, which we have just proved, is equivalent to $X = \partial/\partial x^1$. ♦

The second use of the equation

$$(Xf)(p) = \lim_{h \rightarrow 0} \frac{1}{h} [f(\phi_h(p)) - f(p)]$$

is more comprehensive. The fact that Xf can be defined totally in terms of the diffeomorphisms ϕ_h suggests that an action of X on other objects can be

obtained in a similar way. To emphasize the fundamental similarity of these notions, we first introduce the notation

$$L_X f \quad \text{for} \quad Xf.$$

We call $L_X f$ the (Lie) derivative of f with respect to X ; it is another function, whose value at p is denoted variously by $(L_X f)(p) = L_X f(p) = (Xf)(p) = X_p(f)$. Now if ω is a C^∞ covariant vector field, we define a new covariant vector field, the Lie derivative of ω with respect to X , by

$$(L_X \omega)(p) = \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^* \omega)(p) - \omega(p)].$$

This is the limit of certain members of M_p^* . Recall that if $X_p \in M_p$, then

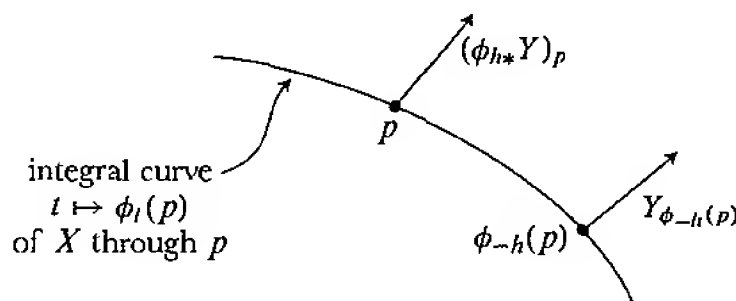
$$(\phi_h^* \omega)(p)(X_p) = \omega(\phi_h(p))(\phi_{h*} X_p).$$

A fairly easy direct argument (Problem 8) shows that this limit always exists, and that the newly defined covariant vector field $L_X \omega$ is C^∞ , but we will soon compute this vector field explicitly in a coordinate system, and these facts will then be obvious.

If Y is another vector field, we can define the Lie derivative of Y with respect to X ,

$$(L_X Y)(p) = \lim_{h \rightarrow 0} \frac{1}{h} [Y_p - (\phi_{h*} Y)_p].$$

The vector field $\phi_{h*} Y$ appearing here is a special case of the vector field $\alpha_* Y$ defined at the beginning of the chapter, for $\alpha: M \rightarrow N$ a diffeomorphism and Y a vector field on M . Thus $(\phi_{h*} Y)_p = \phi_{h*}(Y_{\phi_{-h}(p)})$ is obtained by evaluating Y at $\phi_h^{-1}(p) = \phi_{-h}(p)$, and then moving it back to p by ϕ_{h*} .



The definition of $L_X Y$ can be made to look more closely analogous to $L_X f$ and $L_X \omega$ in the following way. If $\alpha: M \rightarrow N$ is a diffeomorphism and Y is a

vector field on the range N , then a vector field α^*Y on M can be defined by

$$(\alpha^*Y)_p = (\alpha^{-1})_*(Y_{\alpha(p)}).$$

Of course, $\alpha^*(Y)$ is just $(\alpha^{-1})_*Y$. Now notice that

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^*Y)_p - Y_p] &= \lim_{h \rightarrow 0} \frac{Y_p - (\phi_h^*Y)_p}{-h} = \lim_{k \rightarrow 0} \frac{1}{k} [Y_p - (\phi_{-k}^*Y)_p] \\ &= \lim_{k \rightarrow 0} \frac{1}{k} [Y_p - (\phi_k_*Y)_p] = (L_X Y)(p). \end{aligned}$$

Nevertheless, we will stick to the original (equivalent) definition.

We now wish to compute $L_X \omega$ and $L_X Y$ in a coordinate system. The calculation is made a lot easier by first observing

8. PROPOSITION. If $L_X Y_i$ and $L_X \omega_i$ exist for $i = 1, 2$, then

$$(1) \quad L_X (Y_1 + Y_2) = L_X Y_1 + L_X Y_2,$$

$$(2) \quad L_X (\omega_1 + \omega_2) = L_X \omega_1 + L_X \omega_2.$$

If $L_X Y$ and $L_X \omega$ exist, then

$$(3) \quad L_X fY = Xf \cdot Y + f \cdot L_X Y,$$

$$(4) \quad L_X f \cdot \omega = Xf \cdot \omega + f \cdot L_X \omega.$$

Finally, if $\omega(Y)$ denotes the function $p \mapsto \omega(p)(Y_p)$ and $L_X \omega$ and $L_X Y$ exist, then

$$(5) \quad L_X (\omega(Y)) = (L_X \omega)(Y) + \omega(L_X Y).$$

PROOF. (1) and (2) are trivial. The remaining equations are all proved by the same trick, the one used in finding $(fg)'(x)$. We will do number (3) here.

$$\begin{aligned} (L_X fY)_p &= \lim_{h \rightarrow 0} \frac{1}{h} [(fY)_p - (\phi_h^* fY)_p] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(p)Y_p - \phi_{h*}(fY)_{\phi_{-h}(p)}] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(p)Y_p - f(\phi_{-h}(p))\phi_{h*}Y_{\phi_{-h}(p)}] \\ &= \lim_{h \rightarrow 0} f(p) \frac{1}{h} [Y_p - \phi_{h*}Y_{\phi_{-h}(p)}] \\ &\quad + \lim_{h \rightarrow 0} \left[\frac{f(p) - f(\phi_{-h}(p))}{h} \right] \phi_{h*}Y_{\phi_{-h}(p)}. \end{aligned}$$

The first limit is clearly $f(p) \cdot L_X Y(p)$. In the second limit, the term in brackets approaches

$$\lim_{k \rightarrow 0} \frac{f(p) - f(\phi_k(p))}{-k} = Xf(p),$$

while an easy argument shows that $\phi_{h*} Y_{\phi_{-h}(p)} \rightarrow Y_p$. ♦

We are now ready to compute L_X in terms of a coordinate system (x, U) on M . Suppose $X = \sum_{i=1}^n a^i \partial/\partial x^i$. We first compute $L_X(dx^i)$. Recall (Problem 4-1) that if $f: M \rightarrow N$ and y is a coordinate system on N , then

$$f^*(dy^i) = \sum_{j=1}^n \frac{\partial(y^i \circ f)}{\partial x^j} dx^j.$$

We can apply this to ϕ_h^* , where y is x . Then

$$\begin{aligned} L_X(dx^i)(p) &= \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^*)dx^i(p) - dx^i(p)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\sum_{j=1}^n \frac{\partial(x^i \circ \phi_h)}{\partial x^j}(p) dx^j(p) - dx^i(p) \right]. \end{aligned}$$

Now the coefficient of $dx^j(p)$ is

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{1}{h} \left[\frac{\partial(x^i \circ \phi_h)}{\partial x^j} - \delta_j^i \right] &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\frac{\partial(x^i \circ \phi_h)}{\partial x^j}(p) - \frac{\partial(x^i \circ \phi_0)}{\partial x^j}(p) \right] \\ (*) &= \frac{\partial}{\partial x^j} \bigg|_p \lim_{h \rightarrow 0} \frac{1}{h} [(x^i \circ \phi_h) - (x^i \circ \phi_0)] \\ &\quad \{\text{this step will be justified in a moment}\} \\ &= \frac{\partial}{\partial x^j} \bigg|_p X(x^i) = \frac{\partial a^i}{\partial x^j}(p). \end{aligned}$$

To justify (*) we note that the map $A(h, q) = x^i(\phi_h(q))$ is C^∞ from $\mathbb{R} \times M$ to \mathbb{R} ; thus $\partial^2 A / \partial h \partial x^j = \partial^2 A / \partial x^j \partial h$, which is what the interchange of limits amounts to.

It now follows that

$$L_X dx^i = \sum_{j=1}^n \frac{\partial a^i}{\partial x^j} dx^j.$$

We could now use (2) and (4) of Proposition 8 to compute $L_X \omega$ in general, but

we are really interested in computing $L_X Y$. To compute $L_X(\partial/\partial x^i)$ we could imitate the calculations of $L_X dx^i$; but there would be a complication, because ϕ_{h*} on vector fields involves one more composition than ϕ_h^* on covariant vector fields. The trick needed to deal with this complication has already been used to prove (3), (4), and (5) of Proposition 8, and we can now use (5) to get the answer immediately:

$$0 = L_X \delta_j^i = L_X \left[dx^i \left(\frac{\partial}{\partial x^j} \right) \right] = (L_X dx^i) \left(\frac{\partial}{\partial x^j} \right) + dx^i \left(L_X \frac{\partial}{\partial x^j} \right),$$

so

$$dx^i \left(L_X \frac{\partial}{\partial x^j} \right) = - \frac{\partial a^i}{\partial x^j};$$

thus,

$$L_X \frac{\partial}{\partial x^j} = - \sum_{i=1}^n \frac{\partial a^i}{\partial x^j} \frac{\partial}{\partial x^i}.$$

Using (3) we obtain

$$\begin{aligned} L_X \left(b^j \frac{\partial}{\partial x^j} \right) &= L_X b^j \cdot \frac{\partial}{\partial x^j} + b^j L_X \left(\frac{\partial}{\partial x^j} \right) \\ &= \sum_{i=1}^n a^i \frac{\partial b^j}{\partial x^i} \frac{\partial}{\partial x^j} - \sum_{i=1}^n b^j \frac{\partial a^i}{\partial x^j} \frac{\partial}{\partial x^i}. \end{aligned}$$

Summing over j and then interchanging i and j in the second double sum we obtain

$$L_X Y = \sum_{j=1}^n \left(\sum_{i=1}^n a^i \frac{\partial b^j}{\partial x^i} - b^i \frac{\partial a^j}{\partial x^i} \right) \frac{\partial}{\partial x^j}, \quad X = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i}, \quad Y = \sum_{i=1}^n b^i \frac{\partial}{\partial x^i}.$$

This somewhat complicated expression immediately leads to a much simpler coordinate-free expression for $L_X Y$. If $f: M \rightarrow N$ is a C^∞ function, then Yf is a function, so $XYf = X(Yf)$ makes sense. Clearly

$$X(Yf) = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \left(\sum_{j=1}^n b^j \frac{\partial f}{\partial x^j} \right) = \sum_{i,j} a^i \frac{\partial b^j}{\partial x^i} \frac{\partial f}{\partial x^j} + a^i b^j \frac{\partial^2 f}{\partial x^j \partial x^i}.$$

The second partial derivatives which arise here cancel those in the expression for $Y(Xf)$, and we find that

$$L_X Y = XY - YX, \quad \text{also denoted by } [X, Y].$$

Often, $[X, Y]$ (which is called the “bracket” of X and Y) is just defined as $XY - YX$; note that this means

$$[X, Y]_p(f) = X_p(Yf) - Y_p(Xf).$$

A straightforward verification shows that

$$[X, Y]_p(fg) = f(p)[X, Y]_p(g) + g(p)[X, Y]_p(f),$$

so that $[X, Y]_p$ is a derivation at p , and can therefore be considered as a member of M_p .

We are now in a very strange situation. Two vector fields $L_X Y$ and $[X, Y]$ have both been defined independently of any coordinate system, but they have been proved equal using a coordinate system. This sort of thing irks some people to no end. Fortunately, in this case the coordinate-free proof is short, though hardly obvious.

In Chapter 3 we proved a lemma which for the special case of \mathbb{R} says that a C^∞ function $f: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$ with $f(0) = 0$ can be written

$$f(t) = t g(t)$$

for a C^∞ function $g: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$ with $g(0) = f'(0)$, namely

$$g(t) = \int_0^1 f'(st) ds.$$

This has an immediate generalization.

9. LEMMA. If $f: (-\varepsilon, \varepsilon) \times M \rightarrow \mathbb{R}$ is C^∞ and $f(0, p) = 0$ for all $p \in M$, then there is a C^∞ function $g: (-\varepsilon, \varepsilon) \times M \rightarrow \mathbb{R}$ with

$$\begin{aligned} f(t, p) &= t g(t, p) \\ \frac{\partial f}{\partial t}(0, p) &= g(0, p). \end{aligned}$$

PROOF. Define

$$g(t, p) = \int_0^1 \frac{\partial}{\partial s} f(st, p) ds. \quad \spadesuit$$

10. THEOREM. If X and Y are C^∞ vector fields, then

$$L_X Y = [X, Y].$$

PROOF. Let $f: M \rightarrow \mathbb{R}$ be C^∞ . Let X generate ϕ_t , $|t| < \varepsilon$. By Lemma 9 there is a family of C^∞ functions g_t on M such that

$$\begin{aligned} f \circ \phi_t &= f + t g_t \\ g_0 &= Xf. \end{aligned}$$

Then

$$\begin{aligned} (\phi_{h*} Y)_p(f) &= \phi_{h*}(Y_{\phi_{-h}(p)})(f) = Y_{\phi_{-h}(p)}(f \circ \phi_h) \\ &= Y_{\phi_{-h}(p)}(f + h g_h), \end{aligned}$$

so

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{1}{h} [Y_p - (\phi_{h*} Y)_p](f) &= \lim_{h \rightarrow 0} \frac{1}{h} [(Yf)(p) - (Yf)(\phi_{-h}(p))] \\ &\quad - \lim_{h \rightarrow 0} (Yg_h)(\phi_{-h}(p)) \\ &= (L_X Yf)(p) - (Yg_0)(p) \\ &= X_p(Yf) - Y_p(Xf). \quad \spadesuit \end{aligned}$$

The equality $L_X Y = [X, Y] = XY - YX$ reveals certain facts about $L_X Y$ which are by no means obvious from the definition. Clearly

$$[X, Y] = -[Y, X], \quad \text{so} \quad [X, X] = 0.$$

Consequently,

$$L_X Y = -L_Y X, \quad \text{so} \quad L_X X = 0.$$

Since we obviously have $L_X(aY_1 + bY_2) = aL_X Y_1 + bL_X Y_2$, it follows immediately that L is also linear with respect to X :

$$L_{aX_1 + bX_2} Y = aL_{X_1} Y + bL_{X_2} Y.$$

Finally, a straightforward calculation proves the "Jacobi identity":

$$[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0.$$

This equation is capable of two interpretations in terms of Lie derivatives:

(a) $L_X[Y, Z] = [L_X Y, Z] + [Y, L_X Z],$

(b) as operators on C^∞ functions, we have

$$L_{[X, Y]} = L_X \circ L_Y - L_Y \circ L_X \quad (\text{which might be written as } [L_X, L_Y]).$$

Finally, note that $L_X Y$ is linear over constants only, *not* over the C^∞ functions \mathcal{F} . In fact, Proposition 8, or a simple calculation using the definition of $[X, Y]$, shows that

$$[fX, gY] = fg[X, Y] + f(Xg)Y - g(Yf)X.$$

Thus, the bracket operation $[\ , \]$ is *not* a tensor—that is, $[X, Y]_p$ does not depend only on X_p and Y_p (which is not surprising—what can one do to two vectors in a vector space except take linear combinations of them?), but on the vector fields X and Y . In particular, even if $X_p = 0$, it does not necessarily follow that $[X, Y]_p = 0$ —in the formula

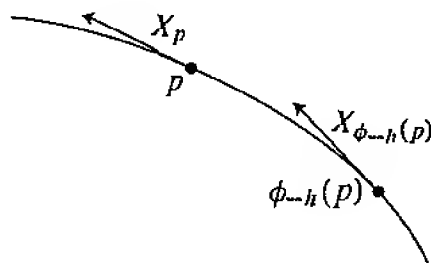
$$[X, Y]_p(f) = X_p(Yf) - Y_p(Xf)$$

the first term $X_p(Yf)$ is zero, but the second may not be, for Xf may have a non-zero derivative in the Y_p direction even though $(Xf)(p) = 0$.

The bracket $[X, Y]$, although not a tensor, pops up in the definition of practically all other tensors, for reasons that will become more and more apparent. Before proceeding to examine its geometric interpretation, we will endeavor to become more at ease with the Lie derivative by taking time out to prove directly from the definition of $L_X Y$ two facts which are obvious from the definition of $[X, Y]$.

(I) $L_X X = 0$.

If X generates ϕ_t , it certainly suffices to show that $(\phi_{h*} X)_p = X_p$ for all h . Recall that $(\phi_{h*} X)_p = \phi_{h*} X_{\phi_{-h}(p)}$. Now $X_{\phi_{-h}(p)}$ is just the tangent vector at



time $t = -h$ to the curve $t \mapsto \phi_t(p)$, and thus the tangent vector, at time $t = 0$, to the curve

$$\gamma(t) = \phi_{t-h}(p).$$

Thus $\phi_{h*} X_{\phi_{-h}(p)}$ is the tangent vector, at time $t = 0$, to the curve

$$\phi_h \circ \gamma(t) = \phi_h(\phi_{t-h}(p)) = \phi_t(p).$$

But this tangent vector is just X_p .

(2) If X_p and Y_p are both 0, then $L_X Y(p) = 0$.

Since $X_p = 0$, the unique integral curve c with $c(0) = p$ and $dc/dt = X(c(t))$ is simply $c(t) = p$ (an integral curve starting at p can never get away; conversely, of course, an integral curve starting at some other point can never get to p). Then $Y_p = 0$ and

$$(\phi_{h*} Y)_p = \phi_{h*} Y_{\phi_{-h}(p)} = \phi_{h*} Y_p = \phi_{h*} 0 = 0,$$

so $L_X Y(p) = 0$.

To develop an interpretation of $[X, Y]$ we first prove two lemmas.

11. LEMMA. Let $\alpha: M \rightarrow N$ be a diffeomorphism and X a vector field on M which generates $\{\phi_t\}$. Then $\alpha_* X$ generates $\{\alpha \circ \phi_t \circ \alpha^{-1}\}$.

PROOF. We have

$$\begin{aligned} (\alpha_* X)_q(f) &= [\alpha_* X_{\alpha^{-1}(q)}](f) \\ &= X_{\alpha^{-1}(q)}(f \circ \alpha) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [(f \circ \alpha)(\phi_h(\alpha^{-1}(q))) - (f \circ \alpha)(\alpha^{-1}(q))] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [f(\alpha \circ \phi_h \circ \alpha^{-1}(q)) - f(q)]. \quad \spadesuit \end{aligned}$$

12. COROLLARY. If $\alpha: M \rightarrow M$, then $\alpha_* X = X$ if and only if $\phi_t \circ \alpha = \alpha \circ \phi_t$ for all t .

13. LEMMA. Let X generate $\{\phi_t\}$ and Y generate $\{\psi_t\}$. Then $[X, Y] = 0$ if and only if $\phi_t \circ \psi_s = \psi_s \circ \phi_t$ for all s, t .

PROOF. If $\phi_t \circ \psi_s = \psi_s \circ \phi_t$ for all s , then $\phi_{t*} Y = Y$ by Corollary 12. If this is true for all t , then clearly $L_X Y = 0$.

Conversely, suppose that $[X, Y] = 0$, so that

$$(*) \quad 0 = \lim_{h \rightarrow 0} \frac{1}{h} [Y_q - (\phi_{h*} Y)_q] \quad \text{for all } q.$$

Given $p \in M$, consider the curve $c: (-\varepsilon, \varepsilon) \rightarrow M_p$ given by

$$c(t) = (\phi_{t*} Y)_p.$$

For the derivative, $c'(t)$, of this map into the vector space M_p we have

$$\begin{aligned}
 c'(t) &= \lim_{h \rightarrow 0} \frac{1}{h} [c(t+h) - c(t)] \\
 &= \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_{[t+h]*} Y)_p - (\phi_{t*} Y)_p] \\
 &= \lim_{h \rightarrow 0} \frac{1}{h} [\phi_{t*} (\phi_{h*} Y)_{\phi_{-t}(p)} - \phi_{t*} Y_{\phi_{-t}(p)}] \\
 &= \phi_{t*} \left\{ \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_{h*} Y)_{\phi_{-t}(p)} - Y_{\phi_{-t}(p)}] \right\} \\
 &= \phi_{t*}(0) \quad \text{using (*) with } q = \phi_{-t}(p) \\
 &= 0.
 \end{aligned}$$

Consequently $c(t) = c(0)$, so $\phi_{t*} Y = Y$. By Corollary 12, $\phi_t \circ \psi_s = \psi_s \circ \phi_t$ for all s, t . ♦

We have already shown that if $X(p) \neq 0$, then there is a coordinate system x with $X = \partial/\partial x^1$. If Y is another vector field, everywhere linearly independent of X , then we might expect to find a coordinate system with

$$(*) \quad X = \frac{\partial}{\partial x^1}, \quad Y = \frac{\partial}{\partial x^2}.$$

However, a short calculation immediately gives the result

$$\left[\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2} \right] = 0,$$

so there is no hope of finding a coordinate system satisfying (*) unless $[X, Y] = 0$. The remarkable fact is that the condition $[X, Y] = 0$ is *sufficient*, as well as necessary, for the existence of the desired coordinate system.

14. THEOREM. If X_1, \dots, X_k are linearly independent C^∞ vector fields in a neighborhood of p , and $[X_\alpha, X_\beta] = 0$ for $1 \leq \alpha, \beta \leq k$, then there is a coordinate system (x, U) around p such that

$$X_\alpha = \frac{\partial}{\partial x^\alpha} \quad \text{on } U, \quad \alpha = 1, \dots, k.$$

PROOF. As in the proof of Theorem 7, we can assume that $M = \mathbb{R}^n$, that $p = 0$, and, by a linear change of coordinates, that

$$X_\alpha(0) = \frac{\partial}{\partial x^\alpha} \Big|_0 \quad \alpha = 1, \dots, k.$$

If X_α generates $\{\phi_t^\alpha\}$, define χ by

$$\chi(a^1, \dots, a^n) = \phi_{a^1}^1(\phi_{a^2}^2(\dots(\phi_{a^k}^k(0, \dots, 0, a^{k+1}, \dots, a^n)) \dots)).$$

As in the proof of Theorem 7, we can compute that

$$\chi_* \left(\frac{\partial}{\partial t^\alpha} \Big|_0 \right) = \begin{cases} X_\alpha(0) = \frac{\partial}{\partial t^\alpha} \Big|_0 & \alpha = 1, \dots, k \\ \frac{\partial}{\partial t^\alpha} \Big|_0 & \alpha = k+1, \dots, n. \end{cases}$$

Thus $x = \chi^{-1}$ can be used as a coordinate system in a neighborhood of $p = 0$. Moreover, just as before we see that

$$X_1 = \frac{\partial}{\partial x^1}.$$

Nothing said so far uses the hypothesis $[X_\alpha, X_\beta] = 0$. To make use of it, we appeal to Lemma 13; it shows that for each α between 1 and k , the map χ can also be written

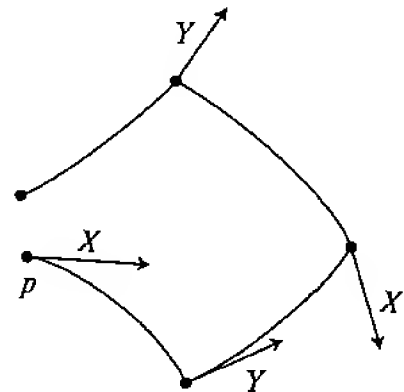
$$\chi(a^1, \dots, a^n) = \phi_{a^\alpha}^\alpha(\phi_{a^1}^1(\dots(0, \dots, 0, a^{k+1}, \dots, a^n) \dots)),$$

and our previous argument then shows that

$$X_\alpha = \frac{\partial}{\partial x^\alpha}. \quad \spadesuit$$

We thus see that the bracket $[X, Y]$ measures, in some sense, the extent to which the integral curves of X and Y can be used to form the “coordinate lines” of a coordinate system. There is a more complicated, more difficult to prove, and less important result, which makes this assertion much more precise. If X and Y are two vector fields in a neighborhood of p , then for sufficiently small h we can

- (1) follow the integral curve of X through p for time h ;
- (2) starting from that point, follow the integral curve of Y for time h ;
- (3) then follow the integral curve of X backwards for time h ;
- (4) then follow the integral curve of Y backwards for time h .



If there happens to be a coordinate system x with $x(p) = 0$ and

$$X = \frac{\partial}{\partial x^1}, \quad Y = \frac{\partial}{\partial x^2},$$

then these steps take us to points with coordinates

$$(1) \quad (h, 0, 0, \dots, 0)$$

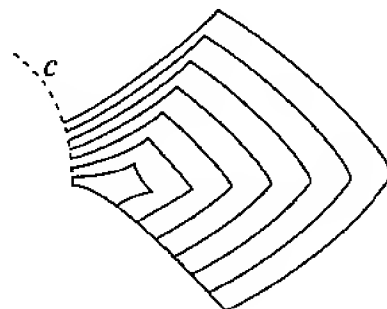
$$(2) \quad (h, h, 0, \dots, 0)$$

$$(3) \quad (0, h, 0, \dots, 0)$$

$$(4) \quad (0, 0, 0, \dots, 0),$$

so that this “parallelogram” is always closed. Even when X and Y are (linearly independent) vector fields with $[X, Y] \neq 0$, the parallelogram is “closed up to first order”. The meaning of this phrase [an extension of the terminology “ $c = \gamma$ up to first order at 0”, which means that $c'(0) = \gamma'(0)$] is the following. Let $c(h)$ be the point which step (4) ends up at,

$$c(h) = \psi_{-h}(\phi_{-h}(\psi_h(\phi_h(0)))).$$



Then the curve c is the constant curve p up to first order, that is,

15. PROPOSITION. $c'(0) = 0$.

PROOF. If we define

$$\alpha_1(t, h) = \psi_t(\phi_h(p))$$

$$\alpha_2(t, h) = \phi_{-t}(\psi_h(\phi_h(p)))$$

$$\alpha_3(t, h) = \psi_{-t}(\phi_{-h}(\psi_h(\phi_h(p)))),$$

then

$$c(t) = \alpha_3(t, t).$$

Moreover,

$$(a) \quad \alpha_2(0, t) = \alpha_1(t, t)$$

$$(b) \quad \alpha_3(0, t) = \alpha_2(t, t)$$

and for any C^∞ function $f: M \rightarrow \mathbb{R}$,

$$\begin{aligned} \text{(c)} \quad & \frac{\partial(f \circ \alpha_1)}{\partial t} = Yf \circ \alpha_1 \\ \text{(d)} \quad & \frac{\partial(f \circ \alpha_2)}{\partial t} = -Xf \circ \alpha_2 \\ \text{(e)} \quad & \frac{\partial(f \circ \alpha_3)}{\partial t} = -Yf \circ \alpha_3 \end{aligned}$$

while

$$\text{(f)} \quad \frac{\partial(f \circ \alpha_1)}{\partial h}(0, h) = Xf(\alpha_1(0, h)).$$

Consequently, repeated use of the chain rule gives

$$\begin{aligned} (f \circ c)'(0) &= D_1(f \circ \alpha_3)(0, 0) + D_2(f \circ \alpha_3)(0, 0) \\ &= D_1(f \circ \alpha_3)(0, 0) \\ &\quad + [D_1(f \circ \alpha_2)(0, 0) + D_2(f \circ \alpha_2)(0, 0)] \quad \text{using (b)} \\ &= D_1(f \circ \alpha_3)(0, 0) + D_1(f \circ \alpha_2)(0, 0) \\ &\quad + [D_1(f \circ \alpha_1)(0, 0) + D_2(f \circ \alpha_1)(0, 0)] \quad \text{using (a)}. \end{aligned}$$

Thus, (c), (d), (e), and (f) give

$$(f \circ c)'(0) = -Yf(p) - Xf(p) + Yf(p) + Xf(p) = 0. \quad \spadesuit$$

Whenever we have a curve $c: (-\varepsilon, \varepsilon) \rightarrow M$ with $c(0) = p$ and $c'(0) = 0 \in M_p$, we can define a new vector $c''(0)$ or $d^2c/dt^2|_0$ by

$$c''(0)(f) = (f \circ c)''(0).$$

A simple calculation shows, *using the assumption* $c'(0) = 0$, that this operator $c''(0)$ is a derivation, $c''(0) \in M_p$. (A more general construction is presented in Problem 17.) It turns out that for the curve c defined previously, the bracket $[X, Y]_p$ is related to this “second order” derivation. Until we get to Lie groups it will not be clear how anyone ever thought of the next theorem. The proof, which ends the chapter, but can easily be skipped, is an horrendous, but clever, calculation. It is followed by an addendum containing some additional important points about differential equations which are used later, and a second addendum concerning linearly independent vector fields in dimension 2.

16. THEOREM. $c''(0) = 2[X, Y]_p$.

PROOF. Using the notation of the previous proof, since $(f \circ c)(t) = (f \circ \alpha_3)(t, t)$ we have

$$(*) \quad (f \circ c)''(0) = D_{1,1}(f \circ \alpha_3)(0, 0) + 2D_{2,1}(f \circ \alpha_3)(0, 0) + D_{2,2}(f \circ \alpha_3)(0, 0).$$

Now

$$\begin{aligned} (1) \quad D_{1,1}(f \circ \alpha_3)(0, 0) &= D_1(-Yf \circ \alpha_3)(0, 0) && \text{by (e)} \\ &= YYf(p) && \text{by (e).} \end{aligned}$$

We also have

$$\begin{aligned} (2) \quad & 2D_{2,1}(f \circ \alpha_3)(0, 0) \\ &= 2D_1(-Yf \circ \alpha_3) && \text{by (e)} \\ &= 2[D_1(Yf \circ \alpha_2)(0, 0) \\ &\quad + D_2(Yf \circ \alpha_2)(0, 0)] && \text{by (b) and the chain rule} \\ &= 2XYf(p) - 2D_2(Yf \circ \alpha_2)(0, 0) && \text{by (d)} \\ &= 2XYf(p) - 2[D_1(Yf \circ \alpha_1)(0, 0) \\ &\quad + D_2(Yf \circ \alpha_1)(0, 0)] && \text{by (a) and the chain rule} \\ &= 2XYf(p) - 2YYf(p) - 2XYf(p) && \text{by (c) and (f).} \end{aligned}$$

Since (b) gives

$$D_2(f \circ \alpha_3)(0, s) = D_1(f \circ \alpha_2)(s, s) + D_2(f \circ \alpha_2)(s, s),$$

we have

$$\begin{aligned} (3) \quad & D_{2,2}(f \circ \alpha_3)(0, 0) = D_{1,1}(f \circ \alpha_2)(0, 0) + 2D_{2,1}(f \circ \alpha_2)(0, 0) \\ &\quad + D_{2,2}(f \circ \alpha_2)(0, 0) \\ &= D_1(-Xf \circ \alpha_2)(0, 0) + 2D_2(-Xf \circ \alpha_2)(0, 0) \\ &\quad + D_{2,2}(f \circ \alpha_2)(0, 0) && \text{by (d)} \\ &= XXf(p) - 2[D_1(Xf \circ \alpha_1)(0, 0) + D_2(Xf \circ \alpha_1)(0, 0)] \\ &\quad + D_{2,2}(f \circ \alpha_2)(0, 0) && \text{by (d) and the chain rule} \\ &= XXf(p) - 2YXf(p) - 2XXf(p) \\ &\quad + D_{2,2}(f \circ \alpha_2)(0, 0) && \text{by (c) and (f).} \end{aligned}$$

Finally, from

$$D_2(f \circ \alpha_2)(0, s) = D_1(f \circ \alpha_1)(s, s) + D_2(f \circ \alpha_2)(s, s) \quad [\text{from (a)}]$$

we have

$$\begin{aligned} (4) \quad D_{2,2}(f \circ \alpha_2)(0, 0) &= D_{1,1}(f \circ \alpha_1)(0, 0) \\ &\quad + 2D_{2,1}(f \circ \alpha_1)(0, 0) + D_{2,2}(f \circ \alpha_1)(0, 0) \\ &= YYf(p) + 2XYf(p) + XXf(p) \\ &\quad \text{by (c) and (f).} \end{aligned}$$

Substituting (1)–(4) in (*) yields the theorem. ❖

ADDENDUM 1

DIFFERENTIAL EQUATIONS

Although we have always solved differential equations

$$\frac{\partial}{\partial t}\alpha(t, x) = f(\alpha(t, x))$$

with the initial condition

$$\alpha(0, x) = x,$$

we could just as well have required, for some t_0 , that

$$\alpha(t_0, x) = x.$$

To prove this, one can replace 0 by t_0 everywhere in the proof of Theorem 2, or else just replace α by $t \mapsto \alpha(t - t_0, x)$.

Another omission in our treatment of differential equations is more glaring: the differential equations $\alpha'(t) = f(\alpha(t))$ do not even include simple equations of the form $\alpha'(t) = g(t)$, let alone equations like $\alpha'(t) = t\alpha(t)$. In general, we would like to solve equations

$$\begin{aligned}\frac{\partial}{\partial t}\alpha(t, x) &= f(t, \alpha(t, x)) \\ \alpha(0, x) &= x,\end{aligned}$$

where $f: (-c, c) \times U \rightarrow \mathbb{R}^n$. One way to do this is to replace $f(\alpha(t, x))$ by $f(t, \alpha(t, x))$ wherever it occurs in the proof. There is also a clever trick. Define

$$\bar{f}: (-c, c) \times U \rightarrow \mathbb{R}^{n+1}$$

by

$$\bar{f}(s, x) = (1, f(s, x)).$$

Then there is a flow $(\bar{\alpha}^1, \bar{\alpha}^2) = \bar{\alpha}: (-b, b) \times W \rightarrow \mathbb{R} \times \mathbb{R}^n$ with

$$\begin{aligned}\frac{\partial}{\partial t}\bar{\alpha}(t, s, x) &= \bar{f}(\bar{\alpha}(t, s, x)) \\ \bar{\alpha}(0, s, x) &= (s, x).\end{aligned}$$

For the first component function $\bar{\alpha}^1$ this means that

$$\begin{aligned}\frac{\partial}{\partial t}\bar{\alpha}^1(t, s, x) &= 1 \\ \bar{\alpha}^1(0, s, x) &= s;\end{aligned}$$

thus

$$\tilde{\alpha}^1(t, s, x) = s + t.$$

For the second component $\tilde{\alpha}^2$ we have

$$\begin{aligned}\frac{\partial}{\partial t}\tilde{\alpha}^2(t, s, x) &= f(\tilde{\alpha}(t, s, x)) \\ &= f(\tilde{\alpha}^1(t, s, x), \tilde{\alpha}^2(t, s, x)) \\ &= f(s + t, \tilde{\alpha}^2(t, s, x)).\end{aligned}$$

Then

$$\beta(t, x) = \tilde{\alpha}^2(t, 0, x)$$

is the desired flow with

$$\begin{aligned}\frac{\partial}{\partial t}\beta(t, x) &= f(t, \beta(t, x)) \\ \beta(0, x) &= x.\end{aligned}$$

Of course, we could also have arranged for $\beta(t_0, x) = x$ (by first finding $\tilde{\alpha}$ with $\tilde{\alpha}(t_0, s, x) = (s, x)$, *not* by considering the curve $t \mapsto \beta(t - t_0, x)$).

Finally, consider the special case of a *linear* differential equation

$$\alpha'(t) = g(t) \cdot \alpha(t),$$

where g is an $n \times n$ matrix-valued function on (a, b) . In this case

$$f(t, x) = g(t) \cdot x.$$

If c is any $n \times n$ (constant) matrix, then

$$(c \cdot \alpha)'(t) = c \cdot \alpha'(t) = g(t) \cdot c \cdot \alpha(t)$$

so $c \cdot \alpha$ is also a solution of the same differential equation. This remark allows us to prove an important property of linear differential equations, distinguishing them from general differential equations $\alpha'(t) = f(t, \alpha(t))$, which may have solutions defined only on a small time interval, even if $f: (a, b) \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is C^∞ .

17. PROPOSITION. If g is a continuous $n \times n$ matrix-valued function on (a, b) , then the solutions of the equation

$$\alpha'(t) = g(t) \cdot \alpha(t)$$

can all be defined on (a, b) .

PROOF. Notice that continuity of g implies that $f(t, x) = g(t) \cdot x$ is locally Lipschitz. So for any $t_0 \in (a, b)$ we can solve the equation, with any given initial condition, in a neighborhood of t_0 . Extend it as far as possible. If the extended solution α is not defined for all t with $t_0 \leq t < b$, let t_1 be the least upper bound of the set of t 's for which it is defined. Pick β with

$$\begin{aligned}\beta'(t) &= g(t) \cdot \beta(t) \quad \text{for } t \text{ near } t_1 \\ \beta(t_1) &\neq 0.\end{aligned}$$

Then $\beta(t^*) \neq 0$ for $t^* < t_1$ close enough to t_1 . Hence there is c with

$$(c \cdot \beta)(t^*) = \alpha(t^*).$$

By uniqueness, $c \cdot \beta$ coincides with α on the interval where they are defined. Thus α may be extended past t_1 as $c \cdot \beta$, a contradiction. Similarly, α must be defined for all t with $a < t \leq t_0$. ♦

ADDENDUM 2

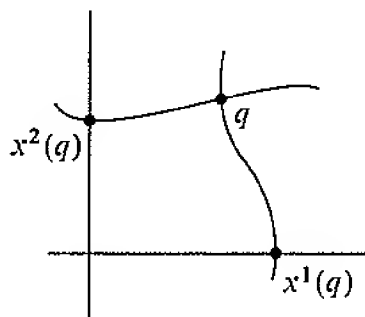
PARAMETER CURVES IN TWO DIMENSIONS

If $f: U \rightarrow M$ is an immersion from an open set $U \subset \mathbb{R}^n$ into an n -dimensional manifold M , the curve $t \mapsto f(a_1, \dots, a_{i-1}, t, a_{i+1}, \dots, a_n)$ is called a *parameter curve* in the i^{th} direction. Given n vector fields X_1, \dots, X_n defined in a neighborhood of $p \in M$ and linearly independent at p , we know that there is usually no immersion $f: U \rightarrow M$ with $p \in f(U)$, whose parameter curves in the i^{th} direction are the integral curves of the X_i —for we might not have $[X_i, X_j] = 0$. However, we might hope to find an immersion f for which the parameter curves in the i^{th} direction lie along the integral curves of the X_i , but have different parameterizations. A simple example (Problem 20) shows that even this modest hope cannot be fulfilled in dimension 3.

On the other hand, in the special case of dimension 2, such an imbedding can be found:

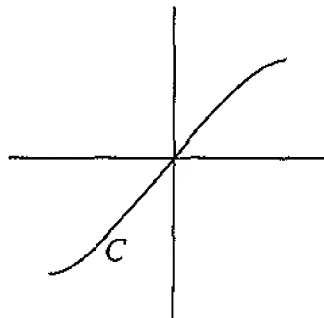
18. PROPOSITION. Let X_1, X_2 be linearly independent vector fields in a neighborhood of a point p in a 2-dimensional manifold M . Then there is an imbedding $f: U \rightarrow M$, where $U \subset \mathbb{R}^2$ is open and $p \in f(U)$, whose i^{th} parameter lines lie along the integral curves of X_i .

PROOF. We can assume that $p = 0 \in \mathbb{R}^2$, and that $X_i(0) = (e_i)_0$. Every point q in a sufficiently small neighborhood of 0 is on a unique integral curve of X_1 through a point $(0, x^2(q))$ —we proved precisely this fact in Theorem 7. Similarly, q is on a unique integral curve of X_2 through a point $(x^1(q), 0)$.



The map $q \mapsto (x^1(q), x^2(q))$ is C^∞ , with Jacobian equal to I at 0 (these facts also follow from the proof of Theorem 7). Its inverse, in a sufficiently small neighborhood of 0, is the required diffeomorphism. ♦

We can always compose f with a map of the form $(x, y) \mapsto (\alpha(x), \beta(y))$ for diffeomorphisms α and β of \mathbb{R} , which gives us considerable flexibility. If, for example, $C \subset \mathbb{R}^2$ is the graph of a monotone function g , then the map



$(x, y) \mapsto (x, g(y))$ takes the diagonal $\{(x, x)\}$ to C . Moreover, for any particular parameterization $c = (c_1, c_2): \mathbb{R} \rightarrow \mathbb{R}^2$ of C , we can further arrange that $c(t)$ maps to $(c(t), c(t))$, by composing with $(x, y) \mapsto (c_1^{-1}(x), y)$. Consequently, we can state

19. PROPOSITION. Let X_1, X_2 be linearly independent vector fields in a neighborhood of a point p in a 2-dimensional manifold M , and let c be a curve in M with $c(0) = p$ and $c'(t)$ never a multiple of X_1 or X_2 . Then there is an imbedding $f: U \rightarrow M$, where $U \subset \mathbb{R}^2$ is open and $p \in f(U)$, whose i^{th} parameter lines lie along the integral curves of X_i , and for which $f(t, t) = c(t)$.

PROBLEMS

1. (a) If $\alpha: M \rightarrow N$ is C^∞ , then $\alpha_*: TM \rightarrow TN$ is C^∞ .
 (b) If $\alpha: M \rightarrow N$ is a diffeomorphism, and X is a C^∞ vector field on M , then α_*X is a C^∞ vector field on N .
 (c) If $\alpha: \mathbb{R} \rightarrow \mathbb{R}$ is $\alpha(t) = t^3$, then there is a C^∞ vector field X on \mathbb{R} such that α_*X is not a C^∞ vector field.
2. Find a nowhere 0 vector field on \mathbb{R} such that all integral curves can be defined only on some interval around 0.
3. Find an example of a complete metric space (M, ρ) and a function $f: M \rightarrow M$ such that $\rho(f(x), f(y)) < \rho(x, y)$ for all $x, y \in M$, but f has no fixed point.
4. Let $f: (-c, c) \times U \times V \rightarrow \mathbb{R}^n$ be C^∞ , where $U, V \subset \mathbb{R}^n$ are open, and let $(x_0, y_0) \in U \times V$. Prove that there is a neighborhood W of (x_0, y_0) and a number $b > 0$ such that for each $(x, y) \in W$ there is a unique $\alpha = \alpha_{(x,y)}: (-b, b) \rightarrow U$ with $\alpha'(t) \in V$ for $t \in (-b, b)$ and

$$\begin{cases} \alpha''(t) = f(t, \alpha(t), \alpha'(t)) \\ \alpha(0) = x \\ \alpha'(0) = y. \end{cases}$$

Moreover, if we write $\alpha_{(x,y)}(t) = \alpha(t, x, y)$, then $\alpha: (-b, b) \times W \rightarrow U$ is C^∞ .
Hint: Consider the system of equations

$$\begin{aligned} \alpha'(t) &= \beta(t) \\ \beta'(t) &= f(t, \alpha(t), \beta(t)). \end{aligned}$$

5. We sometimes have to solve equations “depending on parameters”,

$$\begin{aligned} (*) \quad \frac{\partial}{\partial t} \alpha(t, y, x) &= f(t, y, \alpha(t, y, x)) \\ \alpha(0, y, x) &= x, \end{aligned}$$

where $f: (-c, c) \times V \times U \rightarrow \mathbb{R}^n$, for open $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, and we are solving for $\alpha_{(y,x)}: (-b, b) \rightarrow U$ for each initial condition x and “parameter” y . For example, the equation

$$\begin{aligned} \alpha'(t) &= y\alpha(t) \\ \alpha(0) &= x, \end{aligned}$$

with solution

$$\alpha(t) = xe^{yt},$$

is such a case.

(a) Define

$$\tilde{f}: (-c, c) \times V \times U \rightarrow \mathbb{R}^m \times \mathbb{R}^n$$

by

$$\tilde{f}(t, y, x) = (0, f(t, y, x)).$$

If $(\bar{\alpha}^1, \bar{\alpha}^2) = \bar{\alpha}: (-b, b) \times W \rightarrow \mathbb{R}^m \times \mathbb{R}^n$ is a flow for \tilde{f} in a neighborhood of (y_0, x_0) , so that

$$\begin{aligned} \frac{\partial}{\partial t} \bar{\alpha}(t, y, x) &= \tilde{f}(t, \bar{\alpha}(t, y, x)) \\ \bar{\alpha}(0, y, x) &= (y, x), \end{aligned}$$

show that we can write

$$\bar{\alpha}(t, y, x) = (y, \alpha(t, y, x))$$

for some α , and conclude that α satisfies (*).

(b) Show that equations of the form

$$\begin{aligned} (**) \quad \frac{\partial}{\partial t} \alpha(t, x) &= f(t, x, \alpha(t, x)) \\ \alpha(0, x) &= x \end{aligned}$$

can be reduced to equations of the form (*) (and thus to equations

$$\frac{\partial}{\partial t} \alpha(t, x) = f(\alpha(t, x)),$$

ultimately). [When one proves that a C^k function $f: U \rightarrow \mathbb{R}^n$ has a C^k flow $\alpha: (-b, b) \times W \rightarrow U$, the hard part is to prove that if f is C^1 , then α is differentiable with respect to the arguments in W , and that if the derivative with respect to these arguments is denoted by $D_2\alpha$, then

$$(***) \quad D_1 D_2 \alpha(t, x) = D_2 f(\alpha(t, x)) \cdot D_2 \alpha(t, x)$$

(a result which follows directly from the original equation

$$D_1 \alpha(t, x) = f(\alpha(t, x))$$

if f is C^2 , since $D_1 D_2 = D_2 D_1$). Since (***) is an equation for $D_2 \alpha$ of the form (**), it follows that $D_2 \alpha$ is differentiable if $D_2 f$ is C^1 , i.e., if f is C^2 . Differentiability of class C^k is then proved similarly, by induction.]

6. (a) Consider a linear differential equation

$$\alpha'(t) = g(t)\alpha(t),$$

where $g: \mathbb{R} \rightarrow \mathbb{R}$, so that we are solving for a real-valued function α . Show that all solutions are multiples of

$$\alpha(t) = e^{\int g(t) dt},$$

where $\int g(t) dt$ denotes some function G with $G'(t) = g$ (one can obtain all *positive* multiples simply by changing G). The remainder of this problem investigates the extent to which similar results hold for a system of linear differential equations.

(b) Let $A = (a_{ij})$ be an $n \times n$ matrix, and let $|A|$ denote the maximum of all $|a_{ij}|$. Show that

$$|A + B| \leq |A| + |B|$$

$$|AB| \leq n|A| \cdot |B|.$$

(c) Conclude that the infinite series of $n \times n$ matrices

$$\exp A = e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \frac{A^4}{4!} + \cdots$$

converges absolutely [in the sense that the $(i, j)^{\text{th}}$ entry of the partial sums converge absolutely for each (i, j)] and uniformly in any bounded set.

(d) Show that

$$\exp(TAT^{-1}) = T(\exp A)T^{-1}.$$

(e) If $AB = BA$, then

$$\exp(A + B) = (\exp A)(\exp B).$$

Hint: Write

$$\sum_{p=0}^{2N} \frac{(A+B)^p}{p!} = \left(\sum_{p=0}^N \frac{A^p}{p!} \right) \left(\sum_{p=0}^N \frac{B^p}{p!} \right) + R_N$$

and show that $|R_N| \rightarrow 0$ as $N \rightarrow \infty$.

(f) $(\exp A)(\exp -A) = I$, so $\exp A$ is always invertible.

(g) The map \exp , considered as a map $\exp: \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n^2}$, is clearly differentiable (it is even analytic). Show that

$$\exp'(0)(B) = B \quad (= \exp(0) \cdot B).$$

(Notice that for $|A|$, the usual norm of $A \in \mathbb{R}^{n^2}$, we have $|A| \leq |A| \leq n|A|$.)

(h) Use the limit established in part (g) to show that $\exp'(A)(B) = \exp(A) \cdot B$ if $AB = BA$.

(i) Let $A: \mathbb{R} \rightarrow \mathbb{R}^{n^2}$ be differentiable, and let

$$B(t) = \exp(A(t)).$$

If $B'(t)$ denotes the matrix whose entries are the derivatives of the entries of B , show that

$$B'(t) = A'(t) \cdot \exp(A(t)),$$

provided that $A(t)A'(t) = A'(t)A(t)$. (This is clearly true if $A(s)A(t) = A(t)A(s)$ for all s, t .)

(j) Show that the linear differential equation

$$\alpha'(t) = g(t) \cdot \alpha(t)$$

has the solution

$$\alpha(t) = \exp\left(\int_0^t g(s) ds\right)$$

provided that $g(s)g(t) = g(t)g(s)$ for all s, t . (This certainly happens when $g(t)$ is a constant matrix A , so every system of linear equations with constant coefficients can be solved explicitly—the exponential of $\int_0^t g(s) ds = tA$ can be found by putting A in Jordan canonical form.)

7. Check that if the coordinate system x is $x = \chi^{-1}$, for $\chi: \mathbb{R}^n \rightarrow M$, then $X = \partial/\partial x^1$ is equivalent to $\chi_*(\partial/\partial t^1) = X \circ \chi$.

8. (a) Let M and N be C^∞ manifolds. For a C^∞ function $f: M \times N \rightarrow \mathbb{R}$ and $q \in N$, let $f(\cdot, q)$ denote the function from M to \mathbb{R} defined by

$$p \mapsto f(p, q).$$

If (x, U) is a coordinate system on M , show that the function $\partial f/\partial x^i$, defined by

$$\frac{\partial f}{\partial x^i}(p, q) = \frac{\partial(f(\cdot, q))}{\partial x^i}(p),$$

is a C^∞ function on $M \times N$.

(b) If $\phi: (-\varepsilon, \varepsilon) \times M \rightarrow M$ is a 1-parameter group of diffeomorphisms, show that for every C^∞ function $f: M \rightarrow \mathbb{R}$, the limit

$$\lim_{h \rightarrow 0} \frac{1}{h} [f(\phi_h(p)) - f(p)]$$

exists, and defines a C^∞ function on M .

(c) If $\phi_*: (-\varepsilon, \varepsilon) \times TM \rightarrow TM$ is defined by

$$\phi_*(t, v) = \phi_{t*}(v),$$

show that ϕ_* is C^∞ , and conclude that for every C^∞ vector field X and covariant vector field ω on M , the limit

$$\lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^* \omega)(X_p) - \omega(X_p)]$$

exists and defines a C^∞ function on M .

(d) Treat $L_X Y$ similarly.

9. Give the argument to show that $\phi_{h*} Y_{\phi_{-h}(p)} \rightarrow Y_p$ in the proof of Proposition 8.

10. (a) Prove that

$$\begin{aligned} L_X(f \cdot \omega) &= Xf \cdot \omega + f \cdot L_X \omega \\ L_X[\omega(Y)] &= (L_X \omega)(Y) + \omega(L_X Y). \end{aligned}$$

(b) How would Proposition 8 have to be changed if we had defined $(L_X Y)(p)$ as

$$\lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^* Y)_p - Y_p]?$$

11. (a) Show that

$$\phi^*(df)(Y) = Y(f \circ \phi).$$

(b) Using (a), show directly from the definition of L_X that for $Y \in M_p$,

$$[L_X df(p)](Y_p) = Y_p(L_X f),$$

and conclude that

$$L_X df = d(L_X f).$$

The formula for $L_X dx^i$, derived in the text, is just a special case derived in an unnecessarily clumsy way. In the next part we get a much simpler proof that $L_X Y = [X, Y]$, using the technique which appeared in the proof of Proposition 15.

(c) Let X and Y be vector fields on M , and $f: M \rightarrow \mathbb{R}$ a C^∞ function. If X generates $\{\phi_t\}$, define

$$\alpha(t, h) = Y_{\phi_{-t}(p)}(f \circ \phi_h).$$

Show that

$$D_1\alpha(0,0) = -X_p(Yf)$$

$$D_2\alpha(0,0) = Y_p(Xf).$$

Conclude that for $c(h) = \alpha(h, h)$ we have

$$-c'(0) = L_X Y(p)(f) = [X, Y]_p(f).$$

12. Check the Jacobi identity.

13. On \mathbb{R}^3 let X, Y, Z be the vector fields

$$X = z \frac{\partial}{\partial y} - y \frac{\partial}{\partial z}$$

$$Y = -z \frac{\partial}{\partial x} + x \frac{\partial}{\partial z}$$

$$Z = y \frac{\partial}{\partial x} - x \frac{\partial}{\partial y}.$$

(a) Show that the map

$$aX + bY + cZ \mapsto (a, b, c) \in \mathbb{R}^3$$

is an isomorphism (from a certain set of vector fields to \mathbb{R}^3) and that $[U, V] \mapsto$ the cross-product of the images of U and V .

(b) Show that the flow of $aX + bY + cZ$ is a rotation of \mathbb{R}^3 about some axis through 0.

14. If A is a tensor field of type $\binom{k}{l}$ on N and $\phi: M \rightarrow N$ is a diffeomorphism, we define ϕ^*A on M as follows. If $v_1, \dots, v_k \in M_p$, and $\lambda_1, \dots, \lambda_l \in M_p^*$, then

$$\begin{aligned} [\phi^*A(p)](v_1, \dots, v_k, \lambda_1, \dots, \lambda_l) \\ = A(\phi(p))(\phi_*v_1, \dots, \phi_*v_k, (\phi^{-1})^*\lambda_1, \dots, (\phi^{-1})^*\lambda_l). \end{aligned}$$

(a) Check that under the identification of a vector field [or covariant vector field] with a tensor field of type $\binom{0}{1}$ [or type $\binom{1}{0}$] this agrees with our old ϕ^*Y .

(b) If the vector field X on M generates $\{\phi_t\}$, and A is a tensor field of type $\binom{k}{l}$ on M , we define

$$(L_X A)(p) = \lim_{h \rightarrow 0} \frac{1}{h} [(\phi_h^* A)(p) - A(p)].$$

Show that

$$\begin{aligned} L_X(A + B) &= L_X A + L_X B \\ L_X(A \otimes B) &= (L_X A) \otimes B + A \otimes L_X B \end{aligned}$$

(so that

$$L_X(fA) = X(f)A + fL_X A,$$

in particular).

(c) Show that

$$L_{X_1+X_2} A = L_{X_1} A + L_{X_2} A.$$

Hint: We already know that it is true for A of type $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

(d) Let

$$C: \mathcal{T}_l^k(V) \rightarrow \mathcal{T}_{l-1}^{k-1}(V)$$

be any contraction

$$(CT)(v_1, \dots, v_{k-1}, \lambda_1, \dots, \lambda_{l-1})$$

= contraction of

$$(v, \lambda) \mapsto T(v_1, \dots, v_{\alpha-1}, v, v_{\alpha+1}, \dots, v_{k-1}, \lambda_1, \dots, \lambda_{\beta-1}, \lambda, \lambda_{\beta+1}, \dots, \lambda_{l-1}).$$

Show that

$$L_X(CA) = C(L_X A).$$

(e) Noting that $A(X_1, \dots, X_k, \omega_1, \dots, \omega_l)$ can be obtained by applying contractions repeatedly to $A \otimes X_1 \otimes \dots \otimes X_k \otimes \omega_1 \otimes \dots \otimes \omega_l$, use (d) to show that

$$\begin{aligned} L_X(A(X_1, \dots, X_k, \omega_1, \dots, \omega_l)) \\ &= (L_X A)(X_1, \dots, X_k, \omega_1, \dots, \omega_l) \\ &\quad + \sum_{i=1}^k A(X_1, \dots, L_X X_i, \dots, X_k, \omega_1, \dots, \omega_l) \\ &\quad + \sum_{i=1}^l A(X_1, \dots, X_k, \omega_1, \dots, L_X \omega_i, \dots, \omega_l). \end{aligned}$$

(f) If A has components $A_{i_1 \dots i_k}^{j_1 \dots j_l}$ in a coordinate system x and $X = \sum_{i=1}^n a^i \partial / \partial x^i$, show that the coordinates of $L_X A$ are given by

$$\begin{aligned} (L_X A)_{i_1 \dots i_k}^{j_1 \dots j_l} &= \sum_{i=1}^n a^i \frac{\partial A_{i_1 \dots i_k}^{j_1 \dots j_l}}{\partial x^i} - \sum_{\alpha=1}^k \sum_{j=1}^n A_{i_1 \dots i_k}^{j_1 \dots j_{\alpha-1} j j_{\alpha+1} \dots j_l} \frac{\partial a^{j_{\alpha}}}{\partial x^j} \\ &\quad + \sum_{\alpha=1}^l \sum_{i=1}^n A_{i_1 \dots i_{\alpha-1} i i_{\alpha+1} \dots i_k}^{j_1 \dots j_l} \frac{\partial a^{j_{\alpha}}}{\partial x^i}. \end{aligned}$$

15. Let D be an operator taking the C^∞ functions \mathcal{F} to \mathcal{F} , and the C^∞ vector fields \mathcal{V} to \mathcal{V} , such that $D: \mathcal{F} \rightarrow \mathcal{F}$ and $D: \mathcal{V} \rightarrow \mathcal{V}$ are linear over \mathbb{R} and

$$D(fY) = f \cdot DY + Df \cdot Y.$$

(a) Show that D has a unique extension to an operator taking tensor fields of type $\binom{k}{l}$ to themselves, such that

- (1) D is linear over \mathbb{R}
- (2) $D(A \otimes B) = DA \otimes B + A \otimes DB$
- (3) for any contraction C , $DC = CD$.

If we take $Df = Xf$ and $DY = L_X Y$, then this unique extension is L_X .

(b) Let A be a tensor field of type $\binom{1}{1}$, so that we can consider $A(p) \in \text{End}(M_p)$; then $A(X)$ is a vector field for each vector field X . Show that if we define $D_A f = 0$, $D_A X = A(X)$, then D_A has a unique extension satisfying (1), (2), and (3).

(c) Show that

$$(D_A \omega)(p) = -A(p)^*(\omega(p)).$$

(d) Show that

$$L_f X = f L_X - D_{X \otimes df}.$$

Hint: Check this for functions and vector fields first.

(e) If T is of type $\binom{2}{1}$, show that

$$(D_A T)_k^{ij} = \sum_{\alpha=1}^n T_k^{\alpha j} A_\alpha^i + \sum_{\alpha=1}^n T_k^{i\alpha} A_\alpha^j - \sum_{\alpha=1}^n T_\alpha^{ij} A_k^\alpha.$$

Generalize to tensors of type $\binom{k}{l}$.

16. (a) Let $f: \mathbb{R} \rightarrow \mathbb{R}$ satisfy $f'(0) = 0$. Define $g(t) = f(\sqrt{t})$ for $t \geq 0$. Show that the right-hand derivative

$$g'_+(0) = \lim_{h \rightarrow 0^+} \frac{g(h) - g(0)}{h} = \frac{f''(0)}{2}.$$

(Use Taylor's Theorem.)

(b) Given $c: \mathbb{R} \rightarrow M$ with $c'(0) = 0 \in M_p$, define $\gamma(t) = c(\sqrt{t})$ for $t \geq 0$. Show that the tangent vector $c''(0)$ defined by $c''(0)(f) = (f \circ c)''(0)$ can also be described by $c''(0) = 2\gamma'(0)$.

17. (a) Let $f: M \rightarrow \mathbb{R}$ have p as a critical point, so that $f_{*p} = 0$. Given vectors $X_p, Y_p \in M_p$, choose vector fields \tilde{X}, \tilde{Y} with $\tilde{X}_p = X_p$ and $\tilde{Y}_p = Y_p$. Define

$$f_{**}(X_p, Y_p) = \tilde{X}_p(\tilde{Y}f).$$

Using the fact that $[X, Y]_p(f) = 0$, show that $f_{**}(X_p, Y_p)$ is symmetric, and conclude that it is well-defined.

(b) Show that

$$f_{**}\left(\sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p, \sum_{j=1}^n b^j \frac{\partial}{\partial x^j} \Big|_p\right) = \sum_{i,j=1}^n a^i b^j \frac{\partial^2 f}{\partial x^i \partial x^j}(p).$$

(c) The rank of $(\partial^2 f / \partial x^i \partial x^j(p))$ is independent of the coordinate system.

(d) Let $f: M \rightarrow N$ have p as a critical point. For $X_p, Y_p \in M$ and $g: N \rightarrow \mathbb{R}$ define

$$f_{**}(X, Y)(g) = \tilde{X}_p(\tilde{Y}(g \circ f)).$$

Show that

$$f_{**}: M_p \times M_p \rightarrow N_{f(p)}$$

is a well-defined bilinear map.

(e) If $c: \mathbb{R} \rightarrow M$ has 0 as a critical point, show that

$$c_{**}(0): \mathbb{R}_0 \times \mathbb{R}_0 \rightarrow M_{c(0)}$$

takes $(1_0, 1_0)$ to the tangent vector $c''(0)$ defined by $c''(0)(f) = (f \circ c)''(0)$.

18. Let c be the curve of Theorems 15 and 16. If x is a coordinate system around p with $x(p) = 0$, and

$$[X, Y]_p = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p,$$

show that

$$x^i(c(t)) = a^i t^2 + o(t^2),$$

where $o(t^2)$ denotes a function such that

$$\lim_{t \rightarrow 0} o(t^2)/t^2 = 0.$$

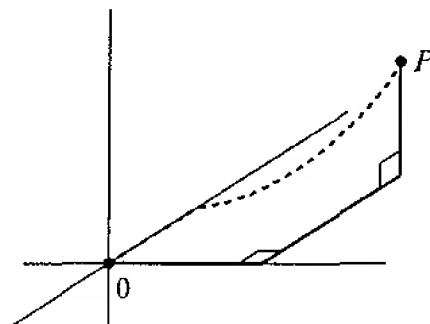
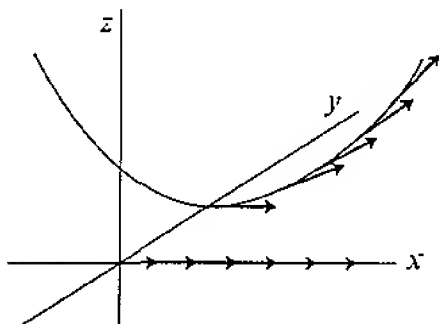
19. (a) If M is compact and 0 is a regular value of $f: M \rightarrow \mathbb{R}$, then there is a neighborhood U of $0 \in \mathbb{R}$ such that $f^{-1}(U)$ is diffeomorphic to $f^{-1}(0) \times U$,

by a diffeomorphism $\phi: f^{-1}(0) \times U \rightarrow f^{-1}(U)$ with $f(\phi(p, t)) = t$. *Hint:* Use Theorem 7 and a partition of unity to construct a vector field X on a neighborhood of $f^{-1}(0)$ such that $f_*X = d/dt$.

(b) More generally, if M is compact and $q \in N$ is a regular value of $f: M \rightarrow N$, then there is a neighborhood U of q and a diffeomorphism $\phi: f^{-1}(q) \times U \rightarrow f^{-1}(U)$ with $f(\phi(p, q')) = q'$.

(c) It follows from (b) that if all points of N are regular values, then $f^{-1}(q_1)$ and $f^{-1}(q_2)$ are diffeomorphic for q_1, q_2 sufficiently close. If f is onto N , does it follow that M is diffeomorphic to $f^{-1}(q) \times N$?

20. In \mathbb{R}^3 , let Y and Z be unit vector fields always pointing along the y - and z -axes, respectively, and let X will be a vector field one of whose integral curves is the x -axis, while certain other integral curves are parabolas in the planes $y = \text{constant}$, as shown in the first part of the figure below. Using the second part of the figure, show that Proposition 18 does not hold in dimension 3.



CHAPTER 6

INTEGRAL MANIFOLDS

PROLOGUE

A mathematician's reputation rests on
the number of bad proofs he has given.
[Pioneer work is clumsy]

A. S. Besicovitch,
quoted in J. E. Littlewood,
A Mathematician's Miscellany

Beauty is the first test: there is no
permanent place in the world for
ugly mathematics.

G. H. Hardy,
A Mathematician's Apology

In the previous chapter, we have seen that the integral curves of a vector field on a manifold M may be definable only for some small time interval, even though the vector field is C^∞ on all of M . We will now vary our question a little, so that global results can be obtained. Instead of a vector field, suppose that for each $p \in M$ we have a 1-dimensional subspace $\Delta_p \subset M_p$. The function Δ is called a **1-dimensional distribution** (this kind of distribution has nothing whatsoever to do with the distributions of analysis, which include such things as the “ δ -function”). Then Δ is spanned by a vector field *locally*; that is, we can choose (in many possible ways) a vector field X such that $0 \neq X_q \in \Delta_q$ for all q in some open set around p . We call Δ a C^∞ distribution if such a vector field X can be chosen to be C^∞ in a neighborhood of each point.

For a 1-dimensional distribution the notion of an integral curve makes no sense, but we define a (1-dimensional) submanifold N of M to be an **integral manifold** of Δ if for every $p \in N$ we have

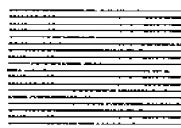
$$i_*(N_p) = \Delta_p \quad \text{where} \quad i: N \rightarrow M \quad \text{is the inclusion map.}$$

For a given $p \in M$, we can always find an integral manifold N of a C^∞ distribution Δ with $p \in N$; we just choose a vector field X with $0 \neq X_q \in \Delta_q$ for q in a neighborhood of p , find an integral curve c of X with initial condition $c(0) = p$, and then forget about the parameterization of c , by defining N to be $\{c(t)\}$. This argument actually shows that for every $p \in M$ there is a coordinate system (x, U) such that for each fixed set of numbers a^2, \dots, a^n , the set

$$\{q \in U : x^2(q) = a^2, \dots, x^n(q) = a^n\}$$

is an integral manifold of Δ on U , and that these are the only integral manifolds in U .

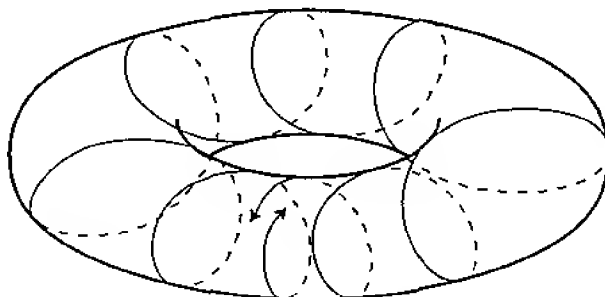
This is still a local result, but because we are dealing with submanifolds, rather than curves with a particular parameterization, we can join overlapping integral submanifolds together. The entire manifold M can be written as a disjoint union of connected integral submanifolds of Δ , which locally look like



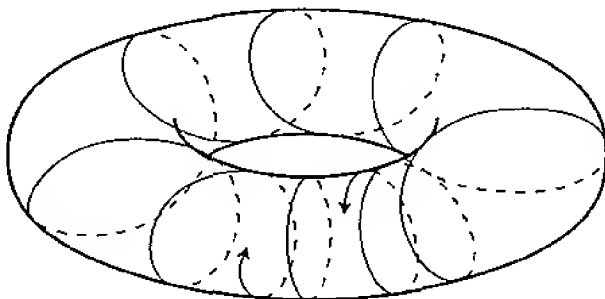
(rather than like



or something even more complicated). For example, there is a distribution on the torus whose integral manifolds all look like the dense 1-dimensional

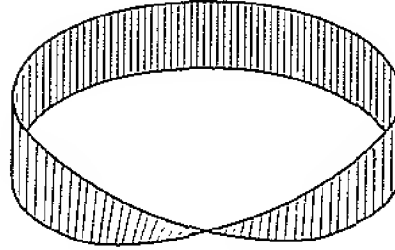


submanifold pictured in Chapter 2. On the other hand, there is a distribution on the torus which has one compact connected integral manifold, and all other



integral manifolds non-compact. It happens that the integral manifolds of these two distributions are also the integral curves for certain vector fields, but on the

Möbius strip there is a distribution which is spanned by a vector field only locally.



We are leaving out the details involved in fitting together these local integral manifolds because we will eventually do this over again in the higher dimensional case. For the moment we will investigate higher dimensional cases only locally.

A k -dimensional distribution on M is a function $p \mapsto \Delta_p$, where $\Delta_p \subset M_p$ is a k -dimensional subspace of M_p . For any $p \in M$ there is a neighborhood U and k vector fields X_1, \dots, X_k such that $X_1(q), \dots, X_k(q)$ are a basis for Δ_q , for each $q \in U$. We call Δ a C^∞ distribution if it is possible to choose C^∞ vector fields X_1, \dots, X_k with this property, in a neighborhood of each point p . A (k -dimensional) submanifold N of M is called an **integral manifold** of Δ if for every $p \in N$ we have

$$i_*(N_p) = \Delta_p \quad \text{where} \quad i: N \rightarrow M \quad \text{is the inclusion map.}$$

Although the definitions given so far all look the same as the 1-dimensional case, the results will look very different. In general, integral manifolds *do not exist*, even locally.

As the simplest example, consider the 2-dimensional distribution Δ in \mathbb{R}^3 for which $\Delta_p = \Delta_{(a,b,c)}$ is spanned by

$$\left. \frac{\partial}{\partial x} \right|_p + b \left. \frac{\partial}{\partial z} \right|_p \quad \text{and} \quad \left. \frac{\partial}{\partial y} \right|_p.$$

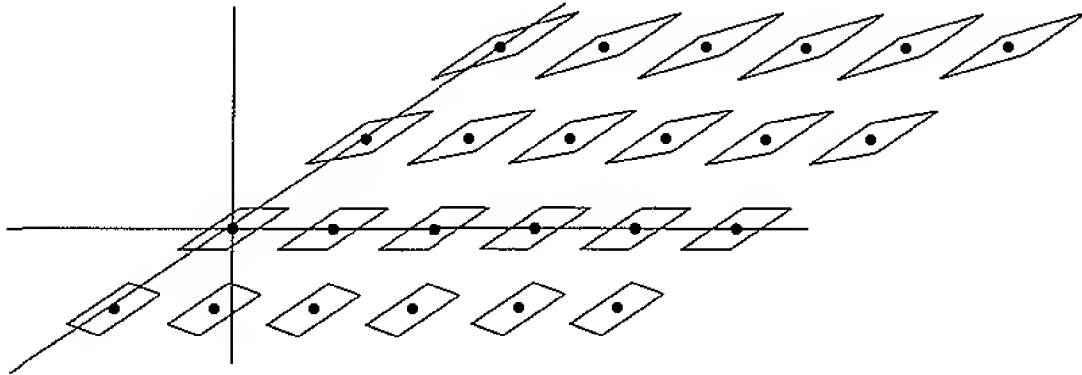
Thus

$$\Delta_p = \left\{ r \left. \frac{\partial}{\partial x} \right|_p + s \left. \frac{\partial}{\partial y} \right|_p + br \left. \frac{\partial}{\partial z} \right|_p : r, s \in \mathbb{R} \right\}.$$

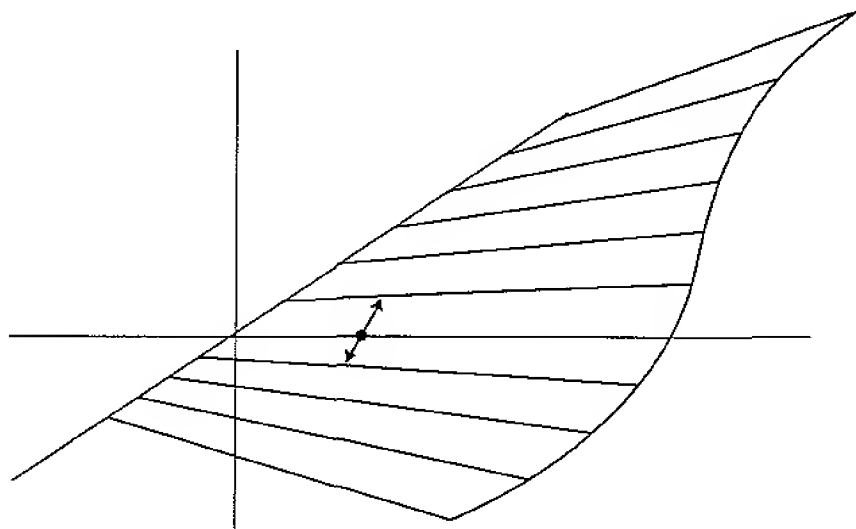
If we identify $T\mathbb{R}^3$ with $\mathbb{R}^3 \times \mathbb{R}^3$, then Δ_p consists of all $(r, s, br)_p$. Thus Δ_p may be pictured as the plane with the equation

$$z - c = b(x - a).$$

The figure below shows Δ_p for points $p = (a, b, 0)$. The plane $\Delta_{(a,b,c)}$ through (a, b, c) is just parallel to the one through $(a, b, 0)$.



If you can picture this distribution, you can probably see that it has no integral manifolds; a proof can be given as follows. Suppose there were an integral manifold N of Δ with $0 \in N$. The intersection of N and $\{(0, y, z)\}$ would be a curve γ in the (y, z) -plane through 0 whose tangent vectors would have to lie in the intersection of $\Delta_{(0,y,z)}$ and the (y, z) -plane. The only such vectors have third component 0 , so γ must be the y -axis. Now consider, for each fixed y_0 , the intersection $N \cap \{(x, y_0, z)\}$. This will be a curve in the plane $\{(x, y_0, z)\}$ through $(0, y_0, 0)$, with all tangent vectors having slope y_0 , so it must be the line $\{(x, y_0, y_0 x)\}$. Our integral manifold would have to look like the following picture. But this submanifold does not work. For example, its tangent space at $(1, 0, 0)$ contains vectors with third component non-zero.



To see in greater detail what is happening here, consider the somewhat more general case where $\Delta_{(a,b,c)} = \Delta_p$ is

$$\Delta_p = \left\{ r \frac{\partial}{\partial x} \Big|_p + s \frac{\partial}{\partial y} \Big|_p + [rf(a,b) + sg(a,b)] \frac{\partial}{\partial z} \Big|_p : r, s \in \mathbb{R} \right\};$$

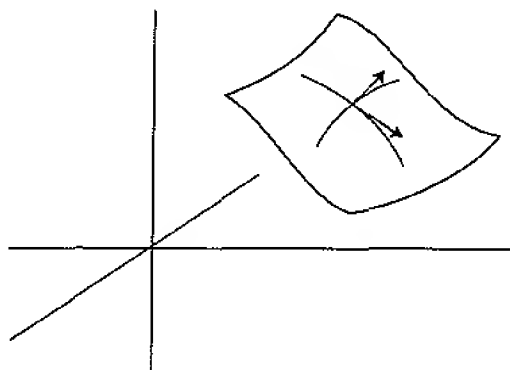
geometrically, Δ_p is the plane with the equation

$$z - c = f(a,b)(x - a) + g(a,b)(y - b).$$

As in the first example, the plane $\Delta_{(a,b,c)}$ through (a,b,c) will be parallel to the one through $(a,b,0)$, since f and g depend only on a and b .

We now ask when the distribution Δ has an integral manifold N through each point. Since Δ_p is never perpendicular to the (x,y) -plane, the submanifold is given locally as the graph of a function:

$$N = \{(x, y, z) : z = \alpha(x, y)\}.$$



Now the tangent space at $p = (a, b, \alpha(a, b))$ is spanned by

$$\frac{\partial}{\partial x} \Big|_p + \frac{\partial \alpha}{\partial x}(a, b) \frac{\partial}{\partial z} \Big|_p,$$

$$\frac{\partial}{\partial y} \Big|_p + \frac{\partial \alpha}{\partial y}(a, b) \frac{\partial}{\partial z} \Big|_p.$$

These tangent vectors are in Δ_p if and only if

$$f(a, b) = \frac{\partial \alpha}{\partial x}(a, b),$$

$$g(a, b) = \frac{\partial \alpha}{\partial y}(a, b).$$

So we need to find a function $\alpha: \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$(*) \quad \frac{\partial \alpha}{\partial x} = f, \quad \frac{\partial \alpha}{\partial y} = g.$$

It is well-known that this is not always possible. By using the equality of mixed partial derivatives, we find a necessary condition on f and g :

$$(**) \quad \frac{\partial f}{\partial y} = \frac{\partial g}{\partial x}.$$

In our previous example,

$$\begin{aligned} f(a, b) &= b, & \frac{\partial f}{\partial y} &= 1, \\ g(a, b) &= 0, & \frac{\partial g}{\partial x} &= 0, \end{aligned}$$

so this necessary condition is not satisfied. It is also well-known that the necessary condition $(**)$ is *sufficient* for the existence of the function α satisfying $(*)$ in a neighborhood of any point.

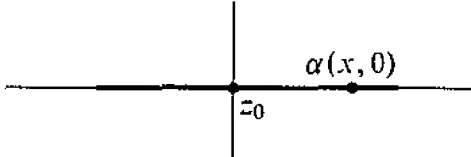
0. PROPOSITION. If $f, g: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfy

$$(**) \quad \frac{\partial f}{\partial y} = \frac{\partial g}{\partial x}$$

in a neighborhood of 0, and $z_0 \in \mathbb{R}$, then there is a function α , defined in a neighborhood of $0 \in \mathbb{R}^2$, such that

$$(*) \quad \begin{aligned} \alpha(0, 0) &= z_0 \\ \frac{\partial \alpha}{\partial x} &= f \\ \frac{\partial \alpha}{\partial y} &= g. \end{aligned}$$

PROOF. We first define $\alpha(x, 0)$ so that $\alpha(0, 0) = z_0$ and

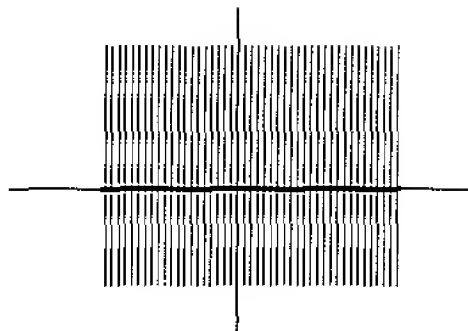
$$(I) \quad \frac{\partial \alpha}{\partial x}(x, 0) = f(x, 0);$$


namely, we define

$$\alpha(x, 0) = z_0 + \int_0^x f(t, 0) dt.$$

Then, for each x , we define $\alpha(x, y)$ so that

$$(2) \quad \frac{\partial \alpha}{\partial y}(x, y) = g(x, y);$$



namely, we define

$$\begin{aligned} \alpha(x, y) &= \alpha(x, 0) + \int_0^y g(x, t) dt \\ &= z_0 + \int_0^x f(t, 0) dt + \int_0^y g(x, t) dt. \end{aligned}$$

This construction does not use (**), and always provide us with an α satisfying (2), $\partial\alpha/\partial y = g$. We claim that if (**) holds, then also $\partial\alpha/\partial x = f$. To prove this, consider, for each fixed x , the function

$$y \mapsto \frac{\partial \alpha}{\partial x}(x, y) - f(x, y).$$

This is 0 for $y = 0$ by (1). To prove that it equals 0 for all y , we just have to show that its derivative is 0. But its derivative at y is

$$\begin{aligned} \frac{\partial^2 \alpha}{\partial y \partial x}(x, y) - \frac{\partial f}{\partial y}(x, y) &= \frac{\partial}{\partial x} \left(\frac{\partial \alpha}{\partial y} \right)(x, y) - \frac{\partial f}{\partial y}(x, y) \\ &= \frac{\partial g}{\partial x}(x, y) - \frac{\partial f}{\partial y}(x, y) \quad \text{by (2)} \\ &= 0 \quad \text{by (**).} \quad \spadesuit \end{aligned}$$

We are now ready to look at essentially the most general case of a 2-dimensional distribution in \mathbb{R}^3 :

$$\Delta_p = \left\{ r \frac{\partial}{\partial x} \Big|_p + s \frac{\partial}{\partial y} \Big|_p + [rf(p) + sg(p)] \frac{\partial}{\partial z} \Big|_p : r, s \in \mathbb{R} \right\},$$

where $f, g: \mathbb{R}^3 \rightarrow \mathbb{R}$. Suppose that

$$N = \{(x, y, z) : z = \alpha(x, y)\}$$

is an integral manifold of Δ . The tangent space of N at $p = (a, b, \alpha(a, b))$ is spanned, once again, by

$$\begin{aligned} \frac{\partial}{\partial x} \Big|_p + \frac{\partial \alpha}{\partial x}(a, b) \frac{\partial}{\partial z} \Big|_p, \\ \frac{\partial}{\partial y} \Big|_p + \frac{\partial \alpha}{\partial y}(a, b) \frac{\partial}{\partial z} \Big|_p. \end{aligned}$$

These tangent vectors are in Δ_p if and only if

$$\begin{aligned} (*) \quad f(a, b, \alpha(a, b)) &= \frac{\partial \alpha}{\partial x}(a, b), \\ g(a, b, \alpha(a, b)) &= \frac{\partial \alpha}{\partial y}(a, b). \end{aligned}$$

In order to obtain necessary conditions for the existence of such a function α , we again use the equality of mixed partial derivatives. Thus (*) and the chain rule imply that

$$\begin{aligned} \frac{\partial^2 \alpha}{\partial y \partial x}(a, b) &= \frac{\partial f}{\partial y}(a, b, \alpha(a, b)) + \frac{\partial f}{\partial z}(a, b, \alpha(a, b)) \cdot \frac{\partial \alpha}{\partial y}(a, b) \\ &\parallel \\ \frac{\partial^2 \alpha}{\partial x \partial y}(a, b) &= \frac{\partial g}{\partial x}(a, b, \alpha(a, b)) + \frac{\partial g}{\partial z}(a, b, \alpha(a, b)) \cdot \frac{\partial \alpha}{\partial y}(a, b). \end{aligned}$$

This condition is not very useful, since it still involves the unknown function α , but we can substitute from (*) to obtain

$$\begin{aligned} \frac{\partial f}{\partial y}(a, b, \alpha(a, b)) + \frac{\partial f}{\partial z}(a, b, \alpha(a, b)) \cdot g(a, b, \alpha(a, b)) \\ = \frac{\partial g}{\partial x}(a, b, \alpha(a, b)) + \frac{\partial g}{\partial z}(a, b, \alpha(a, b)) \cdot f(a, b, \alpha(a, b)). \end{aligned}$$

Now we are looking for conditions which will be satisfied by f and g when there is an integral manifold of Δ *through every point*, which means that for each pair (a, b) these equations must hold no matter what $\alpha(a, b)$ is. Thus we obtain finally the necessary condition

$$(**) \quad \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} \cdot g = \frac{\partial g}{\partial x} + \frac{\partial g}{\partial z} \cdot f.$$

In this more general case, the necessary condition again turns out to be sufficient. In fact, there is no need to restrict ourselves to equations for a single function defined on \mathbb{R}^2 ; we can treat a system of partial differential equations for n functions on \mathbb{R}^m (i.e., a partial differential equation for a function from \mathbb{R}^m to \mathbb{R}^n). In the following theorem, we will use t to denote points in \mathbb{R}^m and x for points in \mathbb{R}^n ; so for a function $f: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^k$ we use

$$\begin{aligned} \frac{\partial f}{\partial t^i} & \quad \text{for } D_i f, \\ \frac{\partial f}{\partial x^i} & \quad \text{for } D_{m+i} f. \end{aligned}$$

1. THEOREM. Let $U \times V \subset \mathbb{R}^m \times \mathbb{R}^n$ be open, where U is a neighborhood of $0 \in \mathbb{R}^m$, and let $f_i: U \times V \rightarrow \mathbb{R}^n$ be C^∞ functions, for $i = 1, \dots, m$. Then for every $x \in V$, there is at most one function

$$\alpha: W \rightarrow V,$$

defined in a neighborhood W of 0 in \mathbb{R}^m , satisfying

$$(*) \quad \begin{aligned} \alpha(0) &= x \\ \frac{\partial \alpha}{\partial t^j}(t) &= f_j(t, \alpha(t)) \quad \text{for all } t \in W. \end{aligned}$$

(More precisely, any two such functions α_1 and α_2 , defined on W_1 and W_2 , agree on the component of $W_1 \cap W_2$ which contains 0 .) Moreover, such a function exists (and is automatically C^∞) in some neighborhood W if and only if there is a neighborhood of $(0, x) \in U \times V$ on which

$$(**) \quad \frac{\partial f_j}{\partial t^i} - \frac{\partial f_i}{\partial t^j} + \sum_{k=1}^n \frac{\partial f_j}{\partial x^k} f_i^k - \sum_{k=1}^n \frac{\partial f_i}{\partial x^k} f_j^k = 0 \quad i, j = 1, \dots, m.$$

PROOF. Uniqueness will be obvious from the proof of existence. Necessity of the conditions (**) is left to the reader as a simple exercise, and we will concern ourselves with proving existence if these conditions do hold. The proof will be like that of Proposition 0, with a different twist at the end.

We first want to define $\alpha(t, 0, \dots, 0)$ so that

$$(I) \quad \begin{aligned} \alpha(0, 0, \dots, 0) &= x \\ \frac{\partial \alpha}{\partial t^1}(t, 0, \dots, 0) &= f_1(t, 0, \dots, 0, \alpha(t, 0, \dots, 0)). \end{aligned}$$

To do this, we consider the ordinary differential equation

$$\begin{aligned}\beta_1(0) &= x \\ \beta_1'(t) &= f_1(t, 0, \dots, 0, \beta_1(t)).\end{aligned}$$

This equation has a unique solution, defined for $|t| < \varepsilon_1$. Define

$$\alpha(t, 0, \dots, 0) = \beta_1(t) \quad |t| < \varepsilon_1.$$

Then (1) holds for $|t| < \varepsilon_1$.

Now for each fixed t^1 with $|t^1| < \varepsilon_1$, consider the equation

$$\begin{aligned}\beta_2(0) &= \alpha(t^1, 0, \dots, 0) \\ \beta_2'(t) &= f_2(t^1, t, 0, \dots, 0, \beta_2(t)).\end{aligned}$$

This has a unique solution for sufficiently small t . At this point the reader must refer back to Theorem 5-2, and verify the following assertion: If we choose ε_1 sufficiently small, then for $|t^1| < \varepsilon_1$ the solutions of the equations for β_2 with the initial conditions $\beta_2(0) = \alpha(t^1, 0, \dots, 0)$ will each be defined for $|t| < \varepsilon_2$ for some $\varepsilon_2 > 0$. We then define

$$\alpha(t^1, t, 0, \dots, 0) = \beta_2(t) \quad |t^1| < \varepsilon_1, |t| < \varepsilon_2.$$

Then

$$\begin{aligned}\alpha(0, 0, 0, \dots, 0) &= x \\ (2) \quad \frac{\partial \alpha}{\partial t^2}(t^1, t, 0, \dots, 0) &= f_2(t^1, t, 0, \dots, 0, \alpha(t^1, t, 0, \dots, 0)) \\ &\quad |t^1| < \varepsilon_1, |t| < \varepsilon_2.\end{aligned}$$

We claim that for each fixed t^1 with $|t^1| < \varepsilon_1$ we also have, for all t with $|t| < \varepsilon_2$,

$$(3) \quad 0 = g(t) = \frac{\partial \alpha}{\partial t^1}(t^1, t, 0, \dots, 0) - f_1(t^1, t, 0, \dots, 0, \alpha(t^1, t, 0, \dots, 0)).$$

Note first that

$$(4) \quad g(0) = 0 \quad \text{by (1).}$$

We now derive an equation for $g'(t)$. In the following, all expressions involving α are to be evaluated at $(t^1, t, 0, \dots, 0)$ and all expressions involving f_i are to be evaluated at $(t^1, t, 0, \dots, 0, \alpha(t^1, t, 0, \dots, 0))$. We have

$$g'(t) = \frac{\partial^2 \alpha}{\partial t^2 \partial t^1} - \frac{\partial f_1}{\partial t^2} - \sum_{k=1}^n \frac{\partial f_1}{\partial x^k} \frac{\partial \alpha^k}{\partial t^2},$$

and thus

$$\begin{aligned}
 (5) \quad g'(t) &= \frac{\partial}{\partial t^1} \left(\frac{\partial \alpha}{\partial t^2} \right) - \frac{\partial f_1}{\partial t^2} - \sum_{k=1}^n \frac{\partial f_1}{\partial x^k} f_2^k && \text{by (2)} \\
 &= \frac{\partial f_2}{\partial t^1} + \sum_{k=1}^n \frac{\partial f_2}{\partial x^k} \frac{\partial \alpha^k}{\partial t^1} - \frac{\partial f_1}{\partial t^2} - \sum_{k=1}^n \frac{\partial f_1}{\partial x^k} f_2^k && \text{by (2) again} \\
 &= \frac{\partial f_2}{\partial t^1} + \sum_{k=1}^n \frac{\partial f_2}{\partial x^k} [g^k(t) + f_1^k] \\
 &\quad - \frac{\partial f_1}{\partial t^2} - \sum_{k=1}^n \frac{\partial f_1}{\partial x^k} f_2^k && \text{by definition, (3)} \\
 &= \sum_{k=1}^n \frac{\partial f_2}{\partial x^k} g^k(t) && \text{by (**).}
 \end{aligned}$$

Now equation (5) is a differential equation with a unique solution for each initial condition. The solution with initial condition $g(0) = 0$, given by (4), is clearly $g(t) = 0$ for all t . So (3) is true.

It is a simple exercise to continue the definition of α until it is eventually defined on $(-\varepsilon_1, \varepsilon_1) \times \cdots \times (-\varepsilon_n, \varepsilon_n)$ and satisfies (*). ♦

Theorem 1 essentially solves for us the problem of deciding which distributions have integral manifolds. Our investigation of the problem so far illustrates one basic fact about theorems in differential geometry:

Many of the fundamental theorems of differential geometry fall into one of two classes. The first kind of theorem says that if one has a certain nice situation (e.g., a distribution with integral submanifolds through every point) then certain other conditions hold; these conditions are obtained by setting mixed partials equal, and are called “integrability conditions”. The second kind of theorem justifies this terminology, by showing that the “integrability conditions” are sufficient for recovering the nice situation.

The remaining parts of our investigation, in which we will essentially begin anew, illustrates an even more important fact about the theorems of differential geometry:

There are always incredibly concise and elegant ways to state the integrability conditions, and prove their sufficiency, without ever even mentioning partial derivatives.

LOCAL THEORY

If $f: M \rightarrow N$ is a C^∞ function, and X and Y are C^∞ vector fields on M and N , respectively, we say that X and Y are f -related if $f_{*p}(X_p) = Y_{f(p)}$ for each $p \in M$. If $g: N \rightarrow \mathbb{R}$ is a C^∞ function, then

$$\begin{aligned} Y_{f(p)}(g) &= f_{*p}X_p(g) \\ &= X_p(g \circ f), \end{aligned}$$

so

$$(Yg) \circ f = X(f \circ g).$$

Conversely, if this is true for all C^∞ functions $g: N \rightarrow \mathbb{R}$, then X and Y are f -related.

Of course, a given vector field X may not be f -related to any vector field Y , nor must a given vector field Y be f -related to any vector field on M . In one case, the latter condition is fulfilled:

2. PROPOSITION. Let $f: M \rightarrow N$ be a C^∞ function such that f is an immersion. If Y is a C^∞ vector field on N with

$$Y_{f(p)} \in f_{*p}(M_p),$$

then there is a unique C^∞ vector field X on M which is f -related to Y .

PROOF. Clearly we must define X_p to be the unique element of M_p with $Y_{f(p)} = f_{*p}X_p$. To prove that X is C^∞ , we use Theorem 2-10(2): there are coordinate systems (x, U) around $p \in M$ and (y, V) around $f(p) \in N$ such that

$$y \circ f \circ x^{-1}(a^1, \dots, a^n) = (a^1, \dots, a^n, 0, \dots, 0).$$

This is easily seen to imply that

$$f_{*p} \left(\frac{\partial}{\partial x^i} \Big|_p \right) = \frac{\partial}{\partial y^i} \Big|_{f(p)}.$$

Thus if

$$Y = \sum_{i=1}^n \alpha^i \frac{\partial}{\partial y^i},$$

where α^i are C^∞ functions, then

$$X = \sum_{i=1}^n \beta^i \frac{\partial}{\partial x^i},$$

where $\alpha^i \circ f = \beta^i$. This implies that the functions β^i are C^∞ (Problem 3). ♦

The most important property of f -relatedness for us is the following:

3. PROPOSITION. If X_i and Y_i are f -related, for $i = 1, 2$, then $[X_1, X_2]$ and $[Y_1, Y_2]$ are f -related.

PROOF. If $g: N \rightarrow \mathbb{R}$ is C^∞ , then

$$(I) \quad (Y_i g) \circ f = X_i(g \circ f) \quad i = 1, 2.$$

So

$$\begin{aligned} \{[Y_1, Y_2]g\} \circ f &= \{Y_1(Y_2 g)\} \circ f - \{Y_2(Y_1 g)\} \circ f \\ &= X_1([Y_2 g] \circ f) - X_2([Y_1 g] \circ f) \\ &\quad \text{by (I), with } g \text{ replaced by } Y_2 g \text{ and } Y_1 g, \text{ respectively} \\ &= X_1(X_2(g \circ f)) - X_2(X_1(g \circ f)) \quad \text{by (I)} \\ &= [X_1, X_2](g \circ f). \quad \diamond \end{aligned}$$

Now consider a k -dimensional distribution Δ . We will say that a vector field X belongs to Δ if $X_p \in \Delta_p$ for all p . Suppose that N is an integral manifold of Δ , and $i: N \rightarrow M$ is the inclusion map. If X and Y are two vector fields which belong to Δ , then for all $p \in N$ there are unique $\bar{X}_p, \bar{Y}_p \in N_p$ such that

$$X_p = i_* \bar{X}_p, \quad Y_p = i_* \bar{Y}_p.$$

In other words, X and \bar{X} are i -related, and Y and \bar{Y} are i -related. Proposition 2 shows that \bar{X} and \bar{Y} are C^∞ vector fields on N , and Proposition 3 then shows that $[\bar{X}, \bar{Y}]$ and $[X, Y]$ are i -related. Thus

$$i_*[\bar{X}, \bar{Y}]_p = [X, Y]_p.$$

Here $[\bar{X}, \bar{Y}]_p \in N_p$; this therefore shows that $[X, Y]_p \in \Delta_p$. Consequently, if there is an integral manifold of Δ through every point p , then $[X, Y]$ also belongs to Δ .

For a moment look back at the distribution Δ in \mathbb{R}^3 given by

$$\Delta_p = \left\{ r \frac{\partial}{\partial x} \Big|_p + s \frac{\partial}{\partial y} \Big|_p + [rf(p) + sg(p)] \frac{\partial}{\partial z} \Big|_p : r, s \in \mathbb{R} \right\}.$$

The vector fields

$$X = \frac{\partial}{\partial x} + f \frac{\partial}{\partial z}$$

$$Y = \frac{\partial}{\partial y} + g \frac{\partial}{\partial z}$$

belong to Δ . Using the formula on page 156, we see that

$$[X, Y] = \left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y} + f \frac{\partial g}{\partial z} - g \frac{\partial f}{\partial z} \right) \cdot \frac{\partial}{\partial z}.$$

This belongs to Δ only when the expression in parentheses is 0, which is precisely the condition for Δ to have an integral manifold through every point.

In general, Δ is called **integrable** if $[X, Y]$ belongs to Δ whenever X and Y belong to Δ . This condition can be checked fairly easily:

4. **PROPOSITION.** If X_1, \dots, X_k span Δ in a neighborhood U of p , then Δ is integrable on U if and only if each $[X_i, X_j]$ is a linear combination

$$[X_i, X_j] = \sum_{\alpha=1}^k C_{ij}^{\alpha} X_{\alpha}$$

for C^{∞} functions C_{ij}^{α} .

PROOF. Such functions clearly exist if Δ is integrable, since $[X_i, X_j]_q \in \Delta_q$, which is spanned by the $X_{\alpha}(q)$. Conversely, suppose such functions exist. If X and Y belong to Δ we can clearly write

$$X = \sum_{i=1}^k f_i X_i$$

$$Y = \sum_{i=1}^k g_i X_i.$$

To prove $[X, Y]$ belongs to Δ , it obviously suffices to treat each $[f_i X_i, g_j X_j]$ separately. Since we have

$$[fX, gY] = fg[X, Y] + f(Xg)Y - g(Yf)X,$$

clearly $[fX, gY]$ belongs to Δ if X, Y and $[X, Y]$ do. ♦

We are now ready for the main theorem. It is equivalent to Theorem 1; in fact, Theorem 1 can be derived from it (Problem 7). But the proof is quite different.

5. **THEOREM (THE FROBENIUS INTEGRABILITY THEOREM; FIRST VERSION).** Let Δ be a C^{∞} integrable k -dimensional distribution on M . For every $p \in M$ there is a coordinate system (x, U) with

$$x(p) = 0$$

$$x(U) = (-\varepsilon, \varepsilon) \times \dots \times (-\varepsilon, \varepsilon),$$

such that for each a^{k+1}, \dots, a^n with all $|a^i| < \varepsilon$, the set

$$\{q \in U : x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\}$$

is an integral manifold of Δ .

Any connected integral manifold of Δ restricted to U is contained in one of these sets.

PROOF. We can clearly assume that we are in \mathbb{R}^n , with $p = 0$. Moreover, we can assume that $\Delta_0 \subset \mathbb{R}^n_0$ is spanned by

$$\left. \frac{\partial}{\partial t^1} \right|_0, \dots, \left. \frac{\partial}{\partial t^k} \right|_0.$$

Let $\pi: \mathbb{R}^n \rightarrow \mathbb{R}^k$ be projection onto the first k factors. Then $\pi_*: \Delta_0 \rightarrow \mathbb{R}^k_0$ is an isomorphism. By continuity, π_* is one-one on Δ_q for q near 0. So near 0, we can choose unique

$$X_1(q), \dots, X_k(q) \in \Delta_q$$

so that

$$\pi_* X_i(q) = \left. \frac{\partial}{\partial t^i} \right|_{\pi(q)} \quad i = 1, \dots, k.$$

Then the vector fields X_i (on a neighborhood of $0 \in \mathbb{R}^n$) and $\partial/\partial t^i$ (on \mathbb{R}^k) are π -related. By Proposition 3,

$$\begin{aligned} \pi_*[X_i, X_j]_q &= \left[\left. \frac{\partial}{\partial t^i} \right|, \left. \frac{\partial}{\partial t^j} \right| \right]_{\pi(q)} \\ &= 0. \end{aligned}$$

But, $[X_i, X_j]_q \in \Delta_q$ by assumption, and π_* is one-one on Δ_q . So $[X_i, X_j] = 0$. By Theorem 5-14, there is a coordinate system x such that

$$X_i = \frac{\partial}{\partial x^i} \quad i = 1, \dots, k.$$

The sets $\{q \in U: x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\}$ are clearly integral manifolds of Δ , since their tangent spaces are spanned by the $\partial/\partial x^i = X_i$ for $i = 1, \dots, k$.

If N is a connected integral manifold of Δ restricted to U , with inclusion map $i: N \rightarrow U$, consider $d(x^m \circ i)$ for $k+1 \leq m \leq n$. For any tangent vector X_q of N_q we have

$$\begin{aligned} d(x^m \circ i)(X_q) &= X_q(x^m \circ i) = i_* X_q(x^m) \\ &= 0, \end{aligned}$$

since $i_* X_q \in \Delta_q$, which is spanned by the $\partial/\partial x^j|_q$ for $j = 1, \dots, k$. Thus $d(x^m \circ i) = 0$, which implies that $x^m \circ i$ is constant on the connected manifold N . ♦

GLOBAL THEORY

In order to express the global results succinctly, we introduce the following terminology.

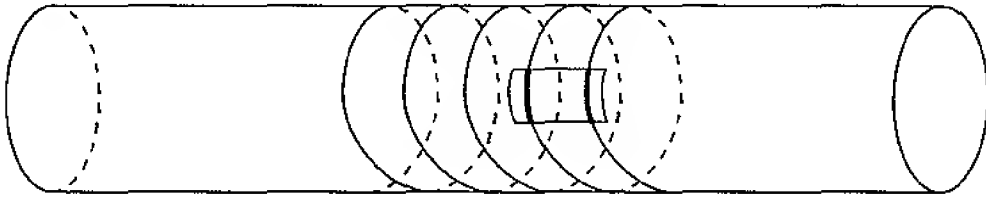
If M is a C^∞ manifold, a (usually disconnected) k -dimensional submanifold N of M is called a **foliation** of M if every point of M is in (some component of) N , and if around every point $p \in M$ there is a coordinate system (x, U) , with

$$x(U) = (-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon),$$

such that the components of $N \cap U$ are the sets of the form

$$\{q \in U : x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\} \quad |a^i| < \varepsilon.$$

Each component of N is called a **folium** or **leaf** of the foliation N . Notice that two distinct components of $N \cap U$ might belong to the same leaf of the foliation.



6. THEOREM. Let Δ be a C^∞ k -dimensional integrable distribution on M . Then M is foliated by an integral manifold of Δ (each component is called a maximal **integral manifold** of Δ).

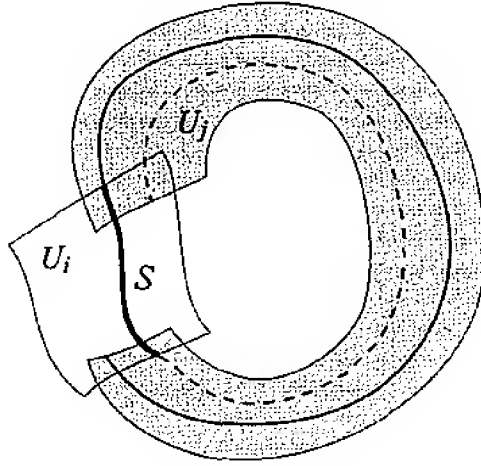
PROOF. Using Theorem 1-2, we see that we can cover M by a sequence of coordinate systems (x_i, U_i) satisfying the conditions of Theorem 5. For such a coordinate system (x, U) , let us call each set

$$\{q \in U : x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\}$$

a *slice* of U .

It is possible for a single slice S of U_i to intersect U_j in more than one slice of U_j , as shown below. But $S \cap U_j$ has at most countably many components,

and each component is contained in a single slice of U_j by Theorem 5, so $S \cap U_j$ is contained in at most countably many slices of U_j .



Given $p \in M$, choose a coordinate system (x_0, U_0) with $p \in U_0$, and let S_0 be the slice of U_0 containing p . A slice S of some U_i will be called *joined* to p if there is a sequence

$$0 = i_0, i_1, \dots, i_l = i$$

and corresponding slices

$$S_0 = S_{i_0}, S_{i_1}, \dots, S_{i_l} = S$$

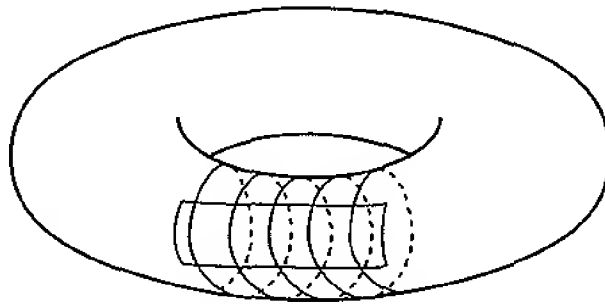
with

$$S_{i_\alpha} \cap S_{i_{\alpha+1}} \neq \emptyset \quad \alpha = 0, \dots, l-1.$$

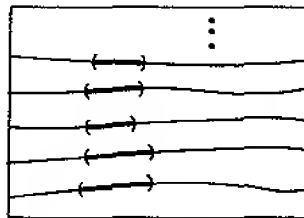
Since there are at most countably many such sequences of slices for each sequence i_0, \dots, i_l , and only countably many such sequences, there are at most countably many slices joined to p . Using Problem 3-1, we see that the union of all such slices is a submanifold of M . For $q \neq p$, the corresponding union is either equal to, or totally disjoint from, the first union. Consequently, M is foliated by the disjoint union of all such submanifolds; this disjoint union is clearly an integral manifold of Δ . ♦

[If we are allowing non-metrizable manifolds, the proof is even easier, since we do not have to find a countable number of coordinate systems for each leaf, and can merely describe the topology of the foliation as the smallest one which makes each slice an open set. In this case, however, the discussion to follow will not be valid—in fact, Appendix A describes a non-paracompact manifold which is foliated by a lower-dimensional *connected* submanifold.]

Notice that if (x, U) is a coordinate system of the sort considered in the proof of the theorem, then infinitely many slices of U may belong to the same folium.



However, *at most countably many* slices can belong to the same folium; otherwise



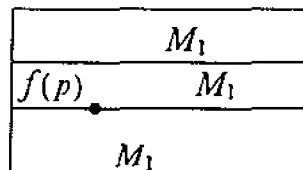
this folium would contain an uncountable disjoint family of open sets. This allows us to apply a proposition from Chapter 2.

7. THEOREM. Let M be a C^∞ manifold, and M_1 a folium of the foliation determined by some distribution Δ . Let P be another C^∞ manifold and $f: P \rightarrow M$ a C^∞ function with $f(P) \subset M_1$. Then f is C^∞ considered as a map into M_1 .

PROOF. According to Proposition 2-11, it suffices to show that f is continuous as a map into M_1 . Given $p \in P$, choose a coordinate system (x, U) around $f(p)$ such that the slices

$$\{q \in U : x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\}$$

are integral manifolds of Δ . Now f is continuous as a map into M , so f takes



some neighborhood W of p into U ; we can choose W to be connected. For $k+1 \leq i \leq n$, if we had $x^i(f(p')) \neq a^i$ for any $p' \in W$, then $x^i \circ f$ would take on all values between a^i and $x^i(f(p'))$, by continuity. This would mean that $f(W)$ contained points of uncountably many slices, contradicting the fact that $f(W) \subset M_1$.

Consequently, $x^i(f(p')) = a^i$ for all $p' \in W$. In other words, $f(W)$ is contained in the single slice of U which contains p . This makes it clear that f is continuous as a map into M_1 . ♦

PROBLEMS

1. (a) Let $\xi = \pi: E \rightarrow B$ be an n -plane bundle, and $\xi' = \pi': E' \rightarrow B$ a k -plane bundle such that $E' \subset E$. If $i: E' \rightarrow E$ is the inclusion map, and $1_B: B \rightarrow B$ the identity map, we say that ξ' is a **subbundle** of ξ if $(i, 1_B)$ is a bundle map. Show that a k -dimensional distribution on M is just a subbundle of TM .

(b) For the case of C^∞ bundles ξ and ξ' over a C^∞ manifold M , define a C^∞ subbundle, and show that a k -dimensional distribution is C^∞ if and only if it is a C^∞ subbundle.

2. (a) In the proof of Theorem 1, check the assertion about choosing ε_1 sufficiently small.

(b) Supply the proof of the uniqueness part of the theorem.

3. (a) In the proof of Proposition 2, show that

$$f_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) = \frac{\partial}{\partial y^i} \Big|_{f(p)}.$$

(b) Complete the proof of Proposition 2 by showing that if

$$Y = \sum_{i=1}^n \alpha^i \frac{\partial}{\partial y^i},$$

so that

$$X = \sum_{i=1}^n \beta^i \frac{\partial}{\partial x^i},$$

with $\alpha^i \circ f = \beta^i$, then the functions β^i are C^∞ .

4. In the proof of Proposition 4, show that the functions C_{ij}^α actually are C^∞ .

5. Let $\Delta_1, \dots, \Delta_h$ be integrable distributions on M , of dimensions d_1, \dots, d_h . Suppose that for each $p \in M$,

$$M_p = (\Delta_1)_p \oplus \cdots \oplus (\Delta_h)_p.$$

Show that there is a coordinate system (x, U) around each point, such that Δ_1 is spanned by $\partial/\partial x^1, \dots, \partial/\partial x^{d_1}$, etc.

6. Prove Theorem 1 from Theorem 5, by considering the distribution Δ in $\mathbb{R}^m \times \mathbb{R}^n$ (with coordinates t, x), defined by

$$\Delta_p = \left\{ \sum_{i=1}^m r^i \frac{\partial}{\partial t^i} \Big|_p + \sum_{k=1}^n \left(\sum_{i=1}^m r^i f_i^k(p) \right) \frac{\partial}{\partial x^k} \Big|_p : r \in \mathbb{R}^m \right\}.$$

Notice that even when the f_j do not depend on x , so that the equations are of the form

$$\frac{\partial \alpha}{\partial t^j}(t) = f_j(t),$$

with the integrability conditions

$$\frac{\partial f_j}{\partial t^i} = \frac{\partial f_i}{\partial t^j},$$

we nevertheless work in $\mathbb{R}^m \times \mathbb{R}^n$, rather than \mathbb{R}^m . This is connected with the classical technique of “introducing new independent variables”.

7. This problem outlines another method of proving Theorem 1, by reducing the partial differential equations to ordinary equations along lines through the origin. A similar technique will be very important in Chapter II.7.

(a) If we want $\alpha(ut) = \beta(u, t)$ for some function $\beta: [0, \varepsilon) \times W \rightarrow V$, show that β must satisfy the equation

$$\begin{aligned} \frac{\partial \beta}{\partial u}(u, t) &= \sum_{j=1}^m t^j \cdot f_j(ut, \beta(u, t)) \\ \beta(0, t) &= x. \end{aligned}$$

We know that we can solve such equations (we need Problem 5-5, since the equation depends on the “parameter” $t \in \mathbb{R}^m$). One has to check that one ε can be picked which works for all $t \in W$.

(b) Show that

$$\beta(u, vt) = \beta(uv, t).$$

(Show that both functions satisfy the same differential equation as functions of u , with the same initial condition.) By shrinking W , we can consequently assume that $\varepsilon = 1$.

(c) Conclude that

$$\frac{\partial \beta}{\partial t^j}(v, t) = v \cdot \frac{\partial \beta}{\partial t^j}(1, vt).$$

(d) Use the integrability condition on f to show that

$$\frac{\partial \beta}{\partial t^j}(v, t) \quad \text{and} \quad v \cdot f_j(vt, \beta(v, t))$$

satisfy the same differential equation, as functions of v . Use (c) to conclude that the two functions are equal.

(e) Define $\alpha(t) = \beta(1, t)$. Noting that $\alpha(vt) = \beta(v, t)$, show that α satisfies the desired equation.

8. This problem is for those who know something about complex analysis. Let $f: \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}$ be complex analytic. If we denote the coordinate functions in $\mathbb{C} \times \mathbb{C}$ by $z_1, z_2 = x_1, y_1, x_2, y_2$, then $f = u + iv$ satisfies the Cauchy-Riemann equations

$$\begin{aligned} \frac{\partial u}{\partial x_i} &= \frac{\partial v}{\partial y_i} \\ \frac{\partial u}{\partial y_i} &= -\frac{\partial v}{\partial x_i} \end{aligned} \quad i = 1, 2.$$

Use Theorem 1 to prove that we can solve the equations

$$\begin{aligned} \frac{\partial \alpha^1}{\partial x} &= u(x, y, \alpha^1(x, y), \alpha^2(x, y)) = \frac{\partial \alpha^2}{\partial y} \\ \frac{\partial \alpha^2}{\partial x} &= v(x, y, \alpha^1(x, y), \alpha^2(x, y)) = -\frac{\partial \alpha^1}{\partial y} \end{aligned}$$

in a neighborhood of $0 \in \mathbb{C}$ (or of any point $z_0 \in \mathbb{C}$), and conclude that the differential equation

$$\phi'(z) = f(z, \phi(z))$$

(in which $'$ denotes the complex derivative) has a solution in a neighborhood of z_0 , with any given initial condition $\phi(z_0) = w_0$.

CHAPTER 7

DIFFERENTIAL FORMS

We turn our attention once more to tensor fields, but we will be concerned with a special kind of tensor field, the discussion of which requires some more algebraic preliminaries.

Let V be an n -dimensional vector space over \mathbb{R} . An element $T \in \mathcal{T}^k(V)$ is called **alternating** if

$$T(v_1, \dots, v_i, \dots, v_j, \dots, v_k) = 0 \quad \text{if } v_i = v_j \ (i \neq j).$$

If T is alternating, then for any v_1, \dots, v_k , we have

$$\begin{aligned} 0 &= T(v_1, \dots, v_i + v_j, \dots, v_i + v_j, \dots, v_k) \\ &= T(v_1, \dots, v_i, \dots, v_i, \dots, v_k) + T(v_1, \dots, v_i, \dots, v_j, \dots, v_k) \\ &\quad + T(v_1, \dots, v_j, \dots, v_i, \dots, v_k) + T(v_1, \dots, v_j, \dots, v_j, \dots, v_k) \\ &= 0 + T(v_1, \dots, v_i, \dots, v_j, \dots, v_k) + T(v_1, \dots, v_j, \dots, v_i, \dots, v_k) + 0. \end{aligned}$$

Therefore, T is **skew-symmetric**:

$$T(v_1, \dots, v_i, \dots, v_j, \dots, v_k) = -T(v_1, \dots, v_j, \dots, v_i, \dots, v_k).$$

Of course, if T is skew-symmetric, then T is also alternating. [This is not true in the special case of a vector space over a field where $1 + 1 = 0$; in this case, skew-symmetry is the same as symmetry, and the condition of being alternating is the stronger one.]

We will denote by $\Omega^k(V)$ the set of all alternating $T \in \mathcal{T}^k(V)$. It is clear that $\Omega^k(V) \subset \mathcal{T}^k(V)$ is a subspace of $\mathcal{T}^k(V)$. Moreover, if $f: V \rightarrow W$ is a linear transformation, then $f^*: \mathcal{T}^k(W) \rightarrow \mathcal{T}^k(V)$ preserves these subspaces— $f^*: \Omega^k(W) \rightarrow \Omega^k(V)$. Notice that $\Omega^1(V) = \mathcal{T}^1(V) = V^*$, so $\Omega^1(V)$ has dimension n . It is also convenient to set $\Omega^0(V) = \mathcal{T}^0(V) = \mathbb{R}$. At the moment it is not clear what the dimension of $\Omega^k(V)$ equals for $k > 1$, but one case is well-known. The most familiar example of an alternating T is the determinant function $\det \in \mathcal{T}^n(\mathbb{R}^n)$, considered as a function of the n rows of a matrix—we shall soon see that this function is, in a certain sense, the most general alternating function. Most discussions of the determinant begin by showing that of any two alternating n -linear functions on \mathbb{R}^n , one is a multiple of the

other; in other words, $\dim \Omega^n(\mathbb{R}^n) \leq 1$. Then one proves $\dim \Omega^n(\mathbb{R}^n) = 1$ by actually constructing the non-zero function \det (it follows, of course, that $\dim \Omega^n(V) = 1$ if V is any n -dimensional vector space). The construction of \det is usually by a messy, explicit formula, which is a special case of the definition to follow.

Let S_k denote the set of all permutations of $\{1, \dots, k\}$; an element $\sigma \in S_k$ is a function $i \mapsto \sigma(i)$. If (v_1, \dots, v_k) is a k -tuple (of any objects) we set

$$\sigma \cdot (v_1, \dots, v_k) = (v_{\sigma(1)}, \dots, v_{\sigma(k)}).$$

This definition has a built-in confusion. On the right side, the first element, for example, is the $\sigma(1)^{\text{st}}$ of the v 's on the left side; if these v 's have indices running in some order *other* than $1, \dots, k$, then the first element on the right is *not* necessarily that v whose index is $\sigma(1)$. The simplest way to figure out something like $\sigma \cdot (v_3, v_2, v_1, \dots)$ is to rename things: $v_3 = w_1, v_2 = w_2, v_1 = w_3, \dots$. Thus warned, we compute

$$\sigma \cdot (\rho \cdot (v_1, \dots, v_k)) = \sigma \cdot (v_{\rho(1)}, \dots, v_{\rho(k)})$$

by setting

$$v_{\rho(1)} = w_1, \dots, v_{\rho(k)} = w_k,$$

so that

$$\begin{aligned} \sigma \cdot (\rho \cdot (v_1, \dots, v_k)) &= \sigma \cdot (w_1, \dots, w_k) \\ &= (w_{\sigma(1)}, \dots, w_{\sigma(k)}) \\ &= (v_{\rho(\sigma(1))}, \dots, v_{\rho(\sigma(k))}) \quad \text{since } w_\alpha = v_{\rho(\alpha)}. \end{aligned}$$

Thus

$$(*) \quad \sigma \cdot (\rho \cdot (v_1, \dots, v_k)) = (\rho\sigma) \cdot (v_1, \dots, v_k).$$

Now for any $T \in \mathcal{T}^k(V)$ we define the "alternation of T "

$$\text{Alt } T = \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot T \circ \sigma,$$

i.e.,

$$\text{Alt } T(v_1, \dots, v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot T(v_{\sigma(1)}, \dots, v_{\sigma(k)}),$$

where $\text{sgn } \sigma$ is $+1$ if σ is an even permutation and -1 if σ is odd.

1. PROPOSITION.

- (1) If $T \in \mathcal{T}^k(V)$, then $\text{Alt}(T) \in \Omega^k(V)$.
- (2) If $\omega \in \Omega^k(V)$, then $\text{Alt } \omega = \omega$.
- (3) If $T \in \mathcal{T}^k(V)$, then $\text{Alt}(\text{Alt}(T)) = \text{Alt}(T)$.

PROOF. Left to the reader (or see pp. 78–79 of *Calculus on Manifolds*). ♦

We now define, for $\omega \in \Omega^k(V)$ and $\eta \in \Omega^l(V)$, an element $\omega \wedge \eta \in \Omega^{k+l}(V)$, the wedge product of ω and η , by

$$\omega \wedge \eta = \frac{(k+l)!}{k!l!} \text{Alt}(\omega \otimes \eta).$$

The funny coefficient is not essential, but it makes some things work out more nicely, as we shall soon see. It is clear that

- (1) \wedge is bilinear:

$$\begin{aligned} (\omega_1 + \omega_2) \wedge \eta &= \omega_1 \wedge \eta + \omega_2 \wedge \eta \\ \omega \wedge (\eta_1 + \eta_2) &= \omega \wedge \eta_1 + \omega \wedge \eta_2 \\ a\omega \wedge \eta &= \omega \wedge a\eta = a(\omega \wedge \eta) \end{aligned}$$

- (2) $f^*(\omega \wedge \eta) = f^*\omega \wedge f^*\eta$.

Moreover, it is easy to see that

- (3) \wedge is “anti-commutative”: $\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega$.

In particular, if k is odd then

$$\omega \wedge \omega = 0.$$

Finally, associativity of \wedge is proved in the following way.

2. THEOREM.

- (1) If $S \in \mathcal{T}^k(V)$ and $T \in \mathcal{T}^l(V)$ and $\text{Alt}(S) = 0$, then

$$\text{Alt}(S \otimes T) = \text{Alt}(T \otimes S) = 0.$$

- (2) $\text{Alt}(\text{Alt}(\omega \otimes \eta) \otimes \theta) = \text{Alt}(\omega \otimes \eta \otimes \theta) = \text{Alt}(\omega \otimes \text{Alt}(\eta \otimes \theta))$.

- (3) If $\omega \in \Omega^k(V)$, $\eta \in \Omega^l(V)$, $\theta \in \Omega^m(V)$, then

$$(\omega \wedge \eta) \wedge \theta = \omega \wedge (\eta \wedge \theta) = \frac{(k+l+m)!}{k!l!m!} \text{Alt}(\omega \otimes \eta \otimes \theta).$$

PROOF. (1) We have

$$\begin{aligned}
 (k+l)! \operatorname{Alt}(S \otimes T)(v_1, \dots, v_{k+l}) \\
 &= \sum_{\sigma \in S_{k+l}} \operatorname{sgn} \sigma \cdot (S \otimes T) \cdot (\sigma \cdot (v_1, \dots, v_{k+l})) \\
 &= \sum_{\sigma \in S_{k+l}} \operatorname{sgn} \sigma \cdot S(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot T(v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}).
 \end{aligned}$$

Now let $G \subset S_{k+l}$ consist of all σ which leave $k+1, \dots, k+l$ fixed. Then

$$\begin{aligned}
 \sum_{\sigma \in G} \operatorname{sgn} \sigma \cdot S(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot T(v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}) \\
 &= \left[\sum_{\sigma' \in S_k} \operatorname{sgn} \sigma' \cdot S(v_{\sigma'(1)}, \dots, v_{\sigma'(k)}) \right] \cdot T(v_{k+1}, \dots, v_{k+l}) \\
 &= 0.
 \end{aligned}$$

Suppose now that $\sigma_0 \notin G$. Let $\sigma_0 G = \{\sigma_0 \sigma' : \sigma' \in G\}$. Then

$$\begin{aligned}
 \sum_{\sigma \in \sigma_0 G} \operatorname{sgn} \sigma \cdot (S \otimes T)(\sigma \cdot (v_1, \dots, v_{k+l})) \\
 = \operatorname{sgn} \sigma_0 \cdot \sum_{\sigma' \in G} \operatorname{sgn} \sigma' \cdot (S \otimes T)(\sigma' \cdot (\sigma_0 \cdot (v_1, \dots, v_{k+l}))) \quad \text{by (*).}
 \end{aligned}$$

We have just shown that this is 0 (since $\sigma_0 \cdot (v_1, \dots, v_{k+l})$ is just some other $(k+l)$ -tuple of vectors). Notice that $G \cap \sigma_0 G = \emptyset$, for if $\sigma \in G \cap \sigma_0 G$, then $\sigma = \sigma_0 \sigma'$ for some $\sigma' \in G$, so $\sigma_0 = \sigma(\sigma')^{-1} \in G$, a contradiction. We can then continue in this way, breaking S_{k+l} up into disjoint subsets, the sum over each being 0. The relation $\operatorname{Alt}(T \otimes S) = 0$ is proved similarly.

(2) Clearly

$$\operatorname{Alt}(\operatorname{Alt}(\eta \otimes \theta) - \eta \otimes \theta) = \operatorname{Alt}(\eta \otimes \theta) - \operatorname{Alt}(\eta \otimes \theta) = 0,$$

so (1) implies that

$$\begin{aligned}
 0 &= \operatorname{Alt}(\omega \otimes [\operatorname{Alt}(\eta \otimes \theta) - \eta \otimes \theta]) \\
 &= \operatorname{Alt}(\omega \otimes \operatorname{Alt}(\eta \otimes \theta)) - \operatorname{Alt}(\omega \otimes \eta \otimes \theta);
 \end{aligned}$$

the other equality is proved similarly.

(3) We have

$$\begin{aligned}
 (\omega \wedge \eta) \wedge \theta &= \frac{(k+l+m)!}{(k+l)!m!} \operatorname{Alt}((\omega \wedge \eta) \otimes \theta) \\
 &= \frac{(k+l+m)!}{(k+l)!m!} \frac{(k+l)!}{k!l!} \operatorname{Alt}(\omega \otimes \eta \otimes \theta).
 \end{aligned}$$

The other equality is proved similarly. ♦

Notice that (2) just states that \wedge is associative even if we had omitted the factor $(k+l)!/k!l!$ in the definition. On the other hand, the factor $1/k!$ in the definition of Alt is essential—without it, we would not have $\text{Alt}(\text{Alt } T) = \text{Alt } T$, and the first equation in the proof of (2) would fail. [If we had defined $\overline{\text{Alt}}$ just like Alt , but without the factor $1/k!$, then \wedge could be defined by

$$\omega \wedge \eta = \frac{1}{k!l!} \overline{\text{Alt}}(\omega \otimes \eta).$$

This makes sense, *even over a field of finite characteristic*, because each term in the sum $\text{Alt}(\omega \otimes \eta)(v_1, \dots, v_{k+l})$ occurs $k!l!$ times (since ω and η are alternating), and $1/k!l!$ can be interpreted as meaning that these $k!l!$ terms are replaced by just one.] The factor $(k+l)!/k!l!$ has been inserted into the definition of \wedge for the following reason. If v_1, \dots, v_n is a basis of V , and ϕ_1, \dots, ϕ_n is the dual basis, then

$$\begin{aligned} \phi_1 \wedge \dots \wedge \phi_n &= \frac{(1 + \dots + 1)!}{1! \dots 1!} \text{Alt}(\phi_1 \otimes \dots \otimes \phi_n) \\ &= \sum_{\sigma \in S_n} \text{sgn } \sigma \cdot (\phi_1 \otimes \dots \otimes \phi_n) \circ \sigma. \end{aligned}$$

In particular,

$$(\phi_1 \wedge \dots \wedge \phi_n)(v_1, \dots, v_n) = 1.$$

(So if v_1, \dots, v_n is the standard basis for \mathbb{R}^n , then $\phi_1 \wedge \dots \wedge \phi_n = \det$.) A basis for $\Omega^k(V)$ can now be described.

3. THEOREM. The set of all

$$\phi_{i_1} \wedge \dots \wedge \phi_{i_k} \quad 1 \leq i_1 < \dots < i_k \leq n$$

is a basis for $\Omega^k(V)$, which therefore has dimension

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

(In particular, $\Omega^k(V) = \{0\}$ for $k > n$.)

PROOF. If $\omega \in \Omega^k(V) \subset \mathcal{T}^k(V)$, we can write

$$\omega = \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} \phi_{i_1} \otimes \dots \otimes \phi_{i_k}.$$

So

$$\omega = \text{Alt}(\omega) = \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} \text{Alt}(\phi_{i_1} \otimes \dots \otimes \phi_{i_k}).$$

Each $\text{Alt}(\phi_{i_1} \otimes \dots \otimes \phi_{i_k})$ is either 0 or $= \pm(1/k!) \phi_{j_1} \wedge \dots \wedge \phi_{j_k}$ for some $j_1 < \dots < j_k$, so the elements $\phi_{j_1} \wedge \dots \wedge \phi_{j_k}$ for $j_1 < \dots < j_k$ span $\Omega^k(V)$. If

$$0 = \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} \phi_{i_1} \wedge \dots \wedge \phi_{i_k},$$

then applying both sides to $(v_{i_1}, \dots, v_{i_k})$ gives $a_{i_1 \dots i_k} = 0$. ♦

4. COROLLARY. If $\omega_1, \dots, \omega_k \in \Omega^1(V)$, then $\omega_1, \dots, \omega_k$ are linearly independent if and only if

$$\omega_1 \wedge \dots \wedge \omega_k \neq 0.$$

PROOF. If $\omega_1, \dots, \omega_k$ are linearly independent, there is a basis $v_1, \dots, v_k, \dots, v_n$ of V such that the dual basis vectors $\phi_1, \dots, \phi_k, \dots, \phi_n$ satisfy $\phi_i = \omega_i$ for $1 \leq i \leq k$. Then $\omega_1 \wedge \dots \wedge \omega_k$ is a basis element of $\Omega^k(V)$, so it is not 0.

On the other hand, if

$$\omega_1 = a_2 \omega_2 + \dots + a_k \omega_k,$$

then

$$\omega_1 \wedge \omega_2 \wedge \dots \wedge \omega_k = (a_2 \omega_2 + \dots + a_k \omega_k) \wedge \omega_2 \wedge \dots \wedge \omega_k = 0. \quad \spadesuit$$

To abbreviate formulas, it is convenient to let I denote a typical “multi-index” (i_1, \dots, i_k) , and let ϕ_I denote $\phi_{i_1} \wedge \dots \wedge \phi_{i_k}$. Then every element of $\Omega^k(V)$ is uniquely expressible as

$$\sum_I a_I \phi_I.$$

Notice that Theorem 3 implies that every $\omega \in \Omega^k(\mathbb{R}^n)$ is a linear combination of the functions

$$(v_1, \dots, v_k) \mapsto \text{determinant of a } k \times k \text{ minor of } \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}.$$

One more simple theorem is in order; before we proceed to apply our construction to manifolds.

5. THEOREM. Let v_1, \dots, v_n be a basis for V , let $\omega \in \Omega^n(V)$, and let

$$w_i = \sum_{j=1}^n \alpha_{ji} v_j \quad i = 1, \dots, n.$$

Then

$$\omega(w_1, \dots, w_n) = \det(\alpha_{ij}) \cdot \omega(v_1, \dots, v_n).$$

PROOF. Define $\eta \in \mathcal{T}^n(\mathbb{R}^n)$ by

$$\eta((a_{11}, \dots, a_{n1}), \dots, (a_{n1}, \dots, a_{nn})) = \omega\left(\sum_{j=1}^n a_{j1} v_j, \dots, \sum_{j=1}^n a_{jn} v_j\right).$$

Then clearly $\eta \in \Omega^n(\mathbb{R}^n)$, so $\eta = c \cdot \det$ for some $c \in \mathbb{R}$, and

$$c = \eta(e_1, \dots, e_n) = \omega(v_1, \dots, v_n). \quad \spadesuit$$

6. COROLLARY. If V is n -dimensional and $0 \neq \omega \in \Omega^n(V)$, then there is a unique orientation μ for V such that

$$[v_1, \dots, v_n] = \mu \quad \text{if and only if} \quad \omega(v_1, \dots, v_n) > 0.$$

With our new algebraic construction at hand, we are ready to apply it to vector bundles. If $\xi = \pi: E \rightarrow B$ is a vector bundle, we obtain a new bundle $\Omega^k(\xi)$ by replacing each fibre $\pi^{-1}(p)$ with $\Omega^k(\pi^{-1}(p))$. A section ω of $\Omega^k(\xi)$ is a function with $\omega(p) \in \Omega^k(\pi^{-1}(p))$ for each $p \in B$. If η is a section of $\Omega^l(\xi)$, then we can define a section $\omega \wedge \eta$ of $\Omega^{k+l}(\xi)$ by $(\omega \wedge \eta)(p) = \omega(p) \wedge \eta(p) \in \Omega^{k+l}(\pi^{-1}(p))$.

In particular, sections of $\Omega^k(TM)$, which are just alternating covariant tensor fields of order k , are called k -forms on M . A 1-form is just a covariant vector field. Since $\Omega^k(TM)$ can obviously be made into a C^∞ vector bundle, we can speak of C^∞ forms; all forms will be understood to be C^∞ forms unless the contrary is explicitly stated. Remember that covariant tensors actually map contravariantly: If $f: M \rightarrow N$ is C^∞ , and ω is a k -form on N , then $f^*\omega$ is a k -form on M . We can also define $\omega_1 + \omega_2$ and $\omega \wedge \eta$. The following properties of k -forms are obvious from the corresponding properties for $\Omega^k(V)$:

$$\begin{aligned} (\omega_1 + \omega_2) \wedge \eta &= \omega_1 \wedge \eta + \omega_2 \wedge \eta \\ \omega \wedge (\eta_1 + \eta_2) &= \omega \wedge \eta_1 + \omega \wedge \eta_2 \\ f\omega \wedge \eta &= \omega \wedge f\eta = f(\omega \wedge \eta) \\ \omega \wedge \eta &= (-1)^{kl} \eta \wedge \omega \\ f^*(\omega \wedge \eta) &= f^*\omega \wedge f^*\eta. \end{aligned}$$

If (x, U) is a coordinate system, then the $dx^i(p)$ are a basis for M_p^* , so the $dx^{i_1}(p) \wedge \cdots \wedge dx^{i_k}(p)$ ($i_1 < \cdots < i_k$) are a basis for $\Omega^k(p)$. Thus every k -form ω can be written uniquely as

$$\omega = \sum_{i_1 < \cdots < i_k} \omega_{i_1 \dots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$$

or, if we denote $dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ by dx^I for the multi-index $I = (i_1, \dots, i_k)$,

$$\omega = \sum_I \omega_I dx^I.$$

The problem of finding the relationship between the ω_I and the functions ω'_I when

$$\omega = \sum_I \omega_I dx^I = \sum_I \omega'_I dy^I$$

is left to the reader (Problem 16), but we will do one special case here.

7. THEOREM. If $f: M \rightarrow N$ is a C^∞ function between n -manifolds, (x, U) is a coordinate system around $p \in M$, and (y, V) a coordinate system around $q = f(p) \in N$, then

$$f^*(g dy^1 \wedge \cdots \wedge dy^n) = (g \circ f) \cdot \det \left(\frac{\partial(y^i \circ f)}{\partial x^j} \right) dx^1 \wedge \cdots \wedge dx^n.$$

PROOF. It suffices to show that

$$f^*(dy^1 \wedge \cdots \wedge dy^n) = \det \left(\frac{\partial(y^i \circ f)}{\partial x^j} \right) dx^1 \wedge \cdots \wedge dx^n.$$

Now, by Problem 4-1,

$$\begin{aligned} f^*(dy^1 \wedge \cdots \wedge dy^n)(p) & \left(\frac{\partial}{\partial x^1} \Big|_p, \dots, \frac{\partial}{\partial x^n} \Big|_p \right) \\ &= dy^1(q) \wedge \cdots \wedge dy^n(q) \left(f_* \frac{\partial}{\partial x^1} \Big|_p, \dots, f_* \frac{\partial}{\partial x^n} \Big|_p \right) \\ &= dy^1(q) \wedge \cdots \wedge dy^n(q) \left(\sum_{i=1}^n \frac{\partial(y^i \circ f)}{\partial x^1}(p) \frac{\partial}{\partial y^i} \Big|_q, \right. \\ & \quad \left. \dots, \sum_{i=1}^n \frac{\partial(y^i \circ f)}{\partial x^n}(p) \frac{\partial}{\partial y^i} \Big|_q \right) \\ &= \det \left(\frac{\partial(y^i \circ f)}{\partial x^j}(p) \right), \quad \text{by Theorem 5. } \spadesuit \end{aligned}$$

8. COROLLARY. If (x, U) and (y, V) are two coordinate systems on M and

$$g \, dy^1 \wedge \cdots \wedge dy^n = h \, dx^1 \wedge \cdots \wedge dx^n,$$

then

$$h = g \cdot \det \left(\frac{\partial y^i}{\partial x^j} \right).$$

PROOF. Apply the theorem with $f = \text{identity map}$. ♦

[This corollary shows that n -forms are the geometric objects corresponding to the “even scalar densities” defined in Problem 4-10.]

If $\xi = \pi : E \rightarrow B$ is an n -plane bundle, then a *nowhere zero* section ω of $\Omega^n(\xi)$ has a special significance: For each $p \in B$, the non-zero $\omega(p) \in \Omega^n(\pi^{-1}(p))$ determines an orientation μ_p of $\pi^{-1}(p)$ by Corollary 6. It is easy to see that the collection of orientations $\{\mu_p\}$ satisfy the “compatibility condition” set forth in Chapter 3, so that $\mu = \{\mu_p\}$ is an orientation of ξ . In particular, if there is a nowhere zero n -form ω on an n -manifold M , then M is orientable (i.e., the bundle TM is orientable). The converse also holds:

9. THEOREM. If a C^∞ manifold M is orientable, then there is an n -form ω on M which is nowhere 0.

PROOF. By Theorem 2-13 and 2-15, we can choose a cover \mathcal{O} of M by a collection of coordinate systems $\{(x, U)\}$, and a partition of unity $\{\phi_U\}$ subordinate to \mathcal{O} . Let μ be an orientation of M . For each (x, U) choose an n -form ω_U on U such that for $v_1, \dots, v_n \in M_p$, $p \in U$ we have

$$\omega_U(v_1, \dots, v_n) > 0 \quad \text{if and only if} \quad [v_1, \dots, v_n] = \mu_p.$$

Now let

$$\omega = \sum_{U \in \mathcal{O}} \phi_U \omega_U.$$

Then ω is a C^∞ n -form. Moreover, for every p , if $v_1, \dots, v_n \in M_p$ satisfy $[v_1, \dots, v_n] = \mu_p$, then *each*

$$(\phi_U \omega_U)(p)(v_1, \dots, v_n) \geq 0,$$

and strict inequality holds for at least one U . Thus $\omega(p) \neq 0$. ♦

Notice that the bundle $\Omega^n(TM)$ is 1-dimensional. We have shown that if M is orientable, then $\Omega^n(TM)$ has a nowhere 0 section, which implies that it is trivial. Conversely, of course, if the bundle $\Omega^n(TM)$ is trivial, then it certainly has a nowhere 0 section, so M is orientable. [Generally, if ξ is a k -plane bundle, then $\Omega^k(\xi)$ is trivial if and only if ξ is orientable, provided that the base space B is “paracompact” (every open cover has a locally-finite refinement).]

Just as $\Omega^0(V)$ has been introduced as another name for \mathbb{R} , a 0-form on M will just mean a function f on M (and $f \wedge \omega$ will just mean $f \cdot \omega$). For every 0-form f we have the 1-form df (recall that $df(X) = X(f)$), which in a coordinate system (x, U) is given by

$$df = \sum_{j=1}^n \frac{\partial f}{\partial x^j} dx^j.$$

If ω is a k -form

$$\omega = \sum_I \omega_I dx^I,$$

then each $d\omega_I$ is a 1-form, and we can define a $(k+1)$ -form $d\omega$, the differential of ω , by

$$\begin{aligned} d\omega &= \sum_I d\omega_I dx^I \\ &= \sum_I \sum_{\alpha=1}^n \frac{\partial \omega_I}{\partial x^\alpha} dx^\alpha \wedge dx^I. \end{aligned}$$

It turns out that this definition does not depend on the coordinate system. This can be proved in several ways. The first way is to use a brute-force computation, comparing the coefficients ω'_I in the expression

$$\omega = \sum_I \omega'_I dx^I$$

with the ω_I .

The second method is a lot sneakier. We begin by finding some properties of $d\omega$ (still defined with respect to this particular coordinate system).

10. PROPOSITION.

(1) $d(\omega_1 + \omega_2) = d\omega_1 + d\omega_2$.

(2) If ω_1 is a k -form, then

$$d(\omega_1 \wedge \omega_2) = d\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2.$$

(3) $d(d\omega) = 0$. Briefly, $d^2 = 0$.

PROOF. (1) is clear. To prove (2) we first note that because of (1) it suffices to consider only

$$\begin{aligned}\omega_1 &= f dx^I \\ \omega_2 &= g dx^J.\end{aligned}$$

Then $\omega_1 \wedge \omega_2 = fg dx^I \wedge dx^J$ and

$$\begin{aligned}d(\omega_1 \wedge \omega_2) &= d(fg) \wedge dx^I \wedge dx^J \\ &= g df \wedge dx^I \wedge dx^J + f dg \wedge dx^I \wedge dx^J \\ &= d\omega_1 \wedge \omega_2 + (-1)^k f dx^I \wedge dg \wedge dx^J \\ &= d\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2.\end{aligned}$$

(3) It clearly suffices to consider only k -forms of the form

$$\omega = f dx^I.$$

Then

$$d\omega = \sum_{\alpha=1}^n \frac{\partial f}{\partial x^\alpha} dx^\alpha \wedge dx^I$$

so

$$d(d\omega) = \sum_{\alpha=1}^n \left(\sum_{\beta=1}^n \frac{\partial^2 f}{\partial x^\beta \partial x^\alpha} dx^\beta \wedge dx^\alpha \wedge dx^I \right).$$

In this sum, the terms

$$\frac{\partial^2 f}{\partial x^\beta \partial x^\alpha} dx^\beta \wedge dx^\alpha \wedge dx^I$$

and

$$\frac{\partial^2 f}{\partial x^\alpha \partial x^\beta} dx^\alpha \wedge dx^\beta \wedge dx^I$$

cancel in pairs. ♦

We next note that these properties characterize d on U .

11. PROPOSITION. Suppose d' takes k -forms on U to $(k+1)$ -forms on U , for all k , and satisfies

- (1) $d'(\omega_1 + \omega_2) = d'\omega_1 + d'\omega_2$.
- (2) $d'(\omega_1 \wedge \omega_2) = d'\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d'\omega_2$.
- (3) $d'(d'f) = 0$.
- (4) $d'f = (\text{the old}) df$.

Then $d' = d$ on U .

PROOF. It is clearly enough to show that $d'\omega = d\omega$ when $\omega = f dx^I$. Now by (2),

$$\begin{aligned} d'(f dx^I) &= d'f \wedge dx^I + f \wedge d'(dx^I) \\ &= df \wedge dx^I + f \wedge d'(dx^I) \quad \text{by (4).} \end{aligned}$$

So it suffices to show that $d'(dx^I) = 0$, where

$$\begin{aligned} dx^I &= dx^{i_1} \wedge \dots \wedge dx^{i_k} \\ &= d'x^{i_1} \wedge \dots \wedge d'x^{i_k} \quad \text{by (4).} \end{aligned}$$

We will use induction on k . Assuming it for $k - 1$ we have

$$\begin{aligned} d'(dx^I) &= d'(d'x^{i_1} \wedge \dots \wedge d'x^{i_k}) \\ &= d'(d'x^{i_1}) \wedge d'x^{i_2} \wedge \dots \wedge d'x^{i_k} \\ &\quad - d'x^{i_1} \wedge d'(d'x^{i_1} \wedge \dots \wedge d'x^{i_k}) \quad \text{by (2)} \\ &= 0 - 0, \quad \text{by (3) and the inductive hypothesis. } \spadesuit \end{aligned}$$

12. COROLLARY. There is a unique operator d from the k -forms on M to the $(k + 1)$ -forms on M , for all k , satisfying

$$\begin{aligned} d(\omega_1 + \omega_2) &= d\omega_1 + d\omega_2 \\ d(\omega_1 \wedge \omega_2) &= d\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2 \\ d^2 &= 0, \end{aligned}$$

and agreeing with the old d on functions.

PROOF. For each coordinate system (x, U) we have a unique d_U defined. Given the form ω , and $p \in M$, pick any U with $p \in U$ and define

$$d\omega(p) = d_U(\omega|U)(p). \quad \spadesuit$$

The third way of proving that the definition of d does not depend on the coordinate system is to give an invariant definition.

13. THEOREM. If ω is a k -form on M , then there is a unique $(k+1)$ -form $d\omega$ on M such that for every set of vector fields X_1, \dots, X_{k+1} we have

$$\begin{aligned}
 (*) \quad d\omega(X_1, \dots, X_{k+1}) &= \sum_{i=1}^{k+1} (-1)^{i+1} X_i(\omega(X_1, \dots, \widehat{X}_i, \dots, X_{k+1})) \\
 &\quad + \sum_{1 \leq i < j \leq k+1} (-1)^{i+j} \omega([X_i, X_j], X_1, \dots, \widehat{X}_i, \dots, \widehat{X}_j, \dots, X_{k+1}) \\
 & (= \Sigma_1 + \Sigma_2, \text{ say})
 \end{aligned}$$

where $\widehat{}$ over X_i indicates that it is omitted. This $(k+1)$ -form agrees with $d\omega$ as defined previously.

PROOF. The operator which takes (X_1, \dots, X_{k+1}) to $\Sigma_1 + \Sigma_2$ is clearly linear over \mathbb{R} . Moreover, it is actually linear over the C^∞ functions \mathcal{F} . In fact, if X_{i_0} is replaced by fX_{i_0} , then Σ_1 becomes

$$f\Sigma_1 + \sum_{i \neq i_0} (-1)^{i+1} (X_i f) \omega(X_1, \dots, \widehat{X}_i, \dots, X_{k+1}),$$

and using the formulas

$$\begin{aligned}
 [fX, Y] &= f[X, Y] - Yf \cdot X \\
 [X, fY] &= f[X, Y] + Xf \cdot Y,
 \end{aligned}$$

it is easily seen that Σ_2 becomes

$$\begin{aligned}
 f\Sigma_2 + \sum_{i < i_0} (-1)^{i+i_0} (X_i f) \omega(X_{i_0}, X_1, \dots, \widehat{X}_i, \dots, \widehat{X}_{i_0}, \dots, X_{k+1}) \\
 - \sum_{i_0 < j} (-1)^{i_0+j} (X_j f) \omega(X_{i_0}, X_1, \dots, \widehat{X}_{i_0}, \dots, \widehat{X}_j, \dots, X_{k+1});
 \end{aligned}$$

a brief inspection then shows that $\Sigma_1 + \Sigma_2$ becomes $f\Sigma_1 + f\Sigma_2$.

Theorem 4-2 shows that there is a unique covariant tensor field $d\omega$ satisfying (*). It is easy to check that $d\omega$ is alternating, so that it is a $(k+1)$ -form.

To compute $d\omega$ in a coordinate system (x, U) it clearly suffices to compute $d(f dx^I)$. Moreover, by renumbering, we might as well assume

$$\omega = f dx^1 \wedge \dots \wedge dx^k.$$

For $d\omega$, as for any form, we have

$$d\omega = \sum_{\alpha_1 < \dots < \alpha_{k+1}} d\omega(\partial/\partial x^{\alpha_1}, \dots, \partial/\partial x^{\alpha_{k+1}}) dx^{\alpha_1} \wedge \dots \wedge dx^{\alpha_{k+1}}.$$

It is clear from (*) that $d\omega(\partial/\partial x^{\alpha_1}, \dots, \partial/\partial x^{\alpha_{k+1}}) = 0$

unless some $(\alpha_1, \dots, \widehat{\alpha_i}, \dots, \alpha_{k+1})$ is a permutation of $(1, \dots, k)$.

Since the α 's are increasing, this happens only if

$$(\alpha_1, \dots, \alpha_{k+1}) = (1, \dots, k, j) \quad j > k,$$

in which case

$$d\omega(\partial/\partial x^{\alpha_1}, \dots, \partial/\partial x^{\alpha_k}, \partial/\partial x^j) = (-1)^k \frac{\partial f}{\partial x^j},$$

so

$$\begin{aligned} d\omega &= \sum_{j>k} (-1)^k \frac{\partial f}{\partial x^j} dx^1 \wedge \dots \wedge dx^k \wedge dx^j \\ &= \sum_{j>k} \frac{\partial f}{\partial x^j} dx^j \wedge dx^1 \wedge \dots \wedge dx^k \\ &= \sum_{j=1}^n \frac{\partial f}{\partial x^j} dx^j \wedge dx^1 \wedge \dots \wedge dx^k, \end{aligned}$$

which is just the old definition. ♦

This is our first real example of an invariant definition of an important tensor, and our first use of Theorem 4-2. We do not find $d\omega(p)(v_1, \dots, v_{k+1})$ directly, but first find $d\omega(X_1, \dots, X_{k+1})$, where X_i are vector fields extending v_i , and then evaluate this function at p . By some sort of magic, this turns out to be independent of the extensions X_1, \dots, X_{k+1} . This may not seem to be much of an improvement over using a coordinate system and checking that the definition is independent of the coordinate system. But we can hardly hope for anything better. After all, although $d\omega(X_1, \dots, X_{k+1})(p)$ does not depend on the values of X_i except at p , it *does* depend on the values of ω at points other than p —this must enter into our formula somehow. One other feature of our definition is common to most invariant definitions of tensors—the presence of a term involving brackets of various vector fields. This term is what makes the operator

linear over the C^∞ functions, but it disappears in computations in a coordinate system.

In the particular case where ω is a 1-form, Theorem 13 gives the following formula.

$$d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y])$$

This enables us to state a second version of Theorem 6-5 (The Frobenius Integrability Theorem) in terms of differential forms. Define the ring $\Omega(M)$ to be the direct sum of the rings of l -forms on M , for all l . If Δ is a k -dimensional distribution on M , then $\mathcal{I}(\Delta) \subset \Omega(M)$ will denote the subring generated by the set of all forms ω with the property that (if ω has degree l)

$$\omega(X_1, \dots, X_l) = 0 \quad \text{whenever } X_1, \dots, X_l \text{ belong to } \Delta.$$

It is clear that $\omega_1 + \omega_2 \in \mathcal{I}(\Delta)$ if $\omega_1, \omega_2 \in \mathcal{I}(\Delta)$, and that $\eta \wedge \omega \in \mathcal{I}(\Delta)$ if $\omega \in \mathcal{I}(\Delta)$ [thus, $\mathcal{I}(\Delta)$ is an ideal in the ring $\Omega(M)$]. Locally, the ideal $\mathcal{I}(\Delta)$ is generated by $n - k$ independent 1-forms $\omega^{k+1}, \dots, \omega^n$. In fact, around any point $p \in M$ we can choose a coordinate system (x, U) so that

$$\left. \frac{\partial}{\partial x^1} \right|_p, \dots, \left. \frac{\partial}{\partial x^k} \right|_p \quad \text{span } \Delta_p.$$

Then

$$dx^1(p) \wedge \dots \wedge dx^k(p) \quad \text{is non-zero on } \Delta_p.$$

By continuity, the same is true for q sufficiently close to p , which by Corollary 4 implies that $dx^1(q), \dots, dx^k(q)$ are linearly independent in Δ_q . Therefore, there are C^∞ functions f_β^α such that

$$dx^\alpha(q) = \sum_{\beta=1}^k f_\beta^\alpha(q) dx^\beta(q) \quad \text{restricted to } \Delta_q \quad \alpha = k+1, \dots, n.$$

We can therefore let

$$\omega^\alpha = dx^\alpha - \sum_{\beta=1}^k f_\beta^\alpha dx^\beta.$$

14. PROPOSITION (THE FROBENIUS INTEGRABILITY THEOREM; SECOND VERSION). A distribution Δ on M is integrable if and only if $d(\mathcal{I}(\Delta)) \subset \mathcal{I}(\Delta)$.

PROOF. Locally we can choose 1-forms $\omega^1, \dots, \omega^n$ which span M_q^* for each q such that $\omega^{k+1}, \dots, \omega^n$ generate $\mathfrak{L}(\Delta)$. Let X_1, \dots, X_n be the vector fields with

$$\omega^i(X_j) = \delta_j^i.$$

Then X_1, \dots, X_k span Δ . So Δ is integrable if and only if there are functions C_{ij}^β with

$$[X_i, X_j] = \sum_{\beta=1}^k C_{ij}^\beta X_\beta \quad i, j = 1, \dots, k.$$

Now

$$d\omega^\alpha(X_i, X_j) = X_i(\omega^\alpha(X_j)) - X_j(\omega^\alpha(X_i)) - \omega^\alpha([X_i, X_j]).$$

For $1 \leq i, j \leq k$ and $\alpha > k$, the first two terms on the right vanish. So $d\omega^\alpha(X_i, X_j) = 0$ if and only if $\omega^\alpha([X_i, X_j]) = 0$. But each $\omega^\alpha([X_i, X_j]) = 0$ if and only if each $[X_i, X_j]$ belongs to Δ (i.e., if Δ is integrable), while each $d\omega^\alpha(X_i, X_j) = 0$ if and only if $d\omega^\alpha \in \mathfrak{L}(\Delta)$. ♦

Notice that since the $\omega^i \wedge \omega^j$ ($i < j$) span $\Omega^2(M_q)$ for each q , we can always write

$$\begin{aligned} d\omega^\alpha &= \sum_{i < j} c_{ij}^\alpha \omega^i \wedge \omega^j \\ &= \sum_j \theta_j^\alpha \wedge \omega^j \quad \text{for certain forms } \theta_j^\alpha. \end{aligned}$$

If $\alpha > k$, and $i_0, j_0 \leq k$ are distinct, we have

$$\begin{aligned} 0 &= d\omega^\alpha(X_{i_0}, X_{j_0}) = \sum_j (\theta_j^\alpha \wedge \omega^j)(X_{i_0}, X_{j_0}) \\ &= \theta_{j_0}^\alpha(X_{i_0}), \end{aligned}$$

so we can write the condition $d(\mathfrak{L}(\Delta)) \subset \mathfrak{L}(\Delta)$ as

$$d\omega^\alpha = \sum_{\beta > k} \theta_\beta^\alpha \wedge \omega^\beta.$$

Once we have introduced a coordinate system (x, U) such that the slices

$$\{q \in U : x^{k+1}(q) = a^{k+1}, \dots, x^n(q) = a^n\}$$

are integral submanifolds of Δ , the forms dx^{k+1}, \dots, dx^n are a basis for $\mathfrak{L}(\Delta)$, so $\omega^{k+1}, \dots, \omega^n$ must be linear combinations of them. We therefore have the following.

15. COROLLARY. If $\omega^{k+1}, \dots, \omega^n$ are linearly independent 1-forms in a neighborhood of $p \in M$, then there are 1-forms θ_β^α ($\alpha, \beta > k$) with

$$d\omega^\alpha = \sum_\beta \theta_\beta^\alpha \wedge \omega^\beta$$

if and only if there are functions f_β^α, g^β ($\alpha, \beta > k$) with

$$\omega^\alpha = \sum_\beta f_\beta^\alpha dg^\beta.$$

Although Theorem 13 warms the heart of many an invariant lover, the cases $k > 1$ will hardly ever be used (a very significant exception occurs in the last chapter of Volume V). Problem 18 gives another invariant definition of $d\omega$, using induction on the degree of ω , which is much simpler. The reader may reflect on the difficulties which would be involved in using the definition of Theorem 13 to prove the following important property of d :

16. PROPOSITION. If $f: M \rightarrow N$ is C^∞ and ω is a k -form on N , then

$$f^*(d\omega) = d(f^*\omega).$$

PROOF. For $p \in M$, let (x, U) be a coordinate system around $f(p)$. We can assume

$$\omega = g dx^{i_1} \wedge \dots \wedge dx^{i_k}.$$

We will use induction on k . For $k = 0$ we have, tracing through some definitions,

$$\begin{aligned} f^*(dg)(X) &= dg(f_*X) = [f_*X](g) = X(g \circ f) \\ &= d(g \circ f)(X) \end{aligned}$$

(and, of course, f^*g is to be interpreted as $g \circ f$). Assuming the formula for $k - 1$, we have

$$\begin{aligned} d(f^*\omega) &= d((f^*g dx^{i_1} \wedge \dots \wedge dx^{i_{k-1}}) \wedge f^*dx^{i_k}) \\ &= d(f^*(g dx^{i_1} \wedge \dots \wedge dx^{i_{k-1}})) \wedge f^*dx^{i_k} + 0 \\ &\quad \text{since } df^*dx^{i_k} = dd(x^{i_k} \circ f) = 0 \\ &= f^*(d(g dx^{i_1} \wedge \dots \wedge dx^{i_{k-1}})) \wedge f^*dx^{i_k} \\ &\quad \text{by the inductive hypothesis} \\ &= f^*(dg \wedge dx^{i_1} \wedge \dots \wedge dx^{i_{k-1}}) \wedge f^*dx^{i_k} \\ &= f^*(dg \wedge dx^{i_1} \wedge \dots \wedge dx^{i_{k-1}} \wedge dx^{i_k}) \\ &= f^*(d\omega). \quad \spadesuit \end{aligned}$$

One property of d qualifies, by the criterion of the previous chapter, as a basic theorem of differential geometry. The relation $d^2 = 0$ is just an elegant way of stating that mixed partial derivatives are equal. There is another set of terminology for stating the same thing. A form ω is called *closed* if $d\omega = 0$ and *exact* if $\omega = d\eta$ for some form η . (The terminology “exact” is classical—differential forms used to be called simply “differentials”; a differential was then called “exact” if it actually was the differential of something. The term “closed” is based on an analogy with chains, which will be discussed in the next chapter.) Since $d^2 = 0$, every exact form is closed. In other words, $d\omega = 0$ is a necessary condition for solving $\omega = d\eta$. If ω is a 1-form

$$\omega = \sum_{i=1}^n \omega_i dx^i,$$

then the condition $d\omega = 0$, i.e.,

$$\frac{\partial \omega_i}{\partial x^j} = \frac{\partial \omega_j}{\partial x^i}$$

is necessary for solving $\omega = df$, i.e.,

$$\frac{\partial f}{\partial x^i} = \omega_i.$$

Now we know from Theorem 6-1 that these conditions are also sufficient. For 2-forms the situation is more complicated, however. If ω is a 2-form on \mathbb{R}^3 ,

$$\omega = A dy \wedge dz - B dx \wedge dz + C dx \wedge dy,$$

then

$$\omega = d(P dx + Q dy + R dz)$$

if and only if

$$\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z} = A$$

$$\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} = B$$

$$\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = C.$$

The necessary condition, $d\omega = 0$, is

$$\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z} = 0.$$

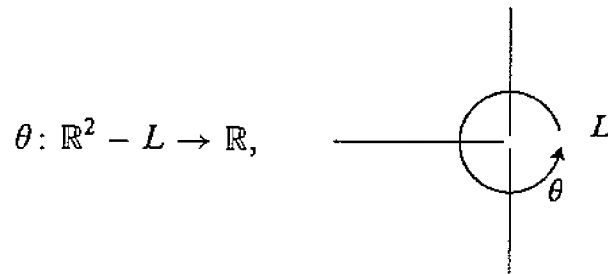
In general, we are dealing with a rather strange collection of partial differential equations (carefully selected so that we can get integrability conditions). It turns out that these necessary conditions are also sufficient: if ω is closed, then it is exact. Like our results about solutions to differential equations, this result is true only locally. The reasons for restricting ourselves to local results are now somewhat different, however. Consider the case of a closed 1-form ω on \mathbb{R}^2 :

$$\omega = f dx + g dy, \quad \text{with} \quad \frac{\partial f}{\partial y} = \frac{\partial g}{\partial x}.$$

We know how to find a function α on *all* of \mathbb{R}^2 with $\omega = d\alpha$, namely

$$\alpha(x, y) = \int_{x_0}^x f(t, y_0) dt + \int_{y_0}^y g(x, t) dt.$$

On the other hand, the situation is very different if ω is defined only on $\mathbb{R}^2 - \{0\}$. Recall that if $L \subset \mathbb{R}^2$ is $[0, \infty) \times \{0\}$, then



$$\theta: \mathbb{R}^2 - L \rightarrow \mathbb{R},$$

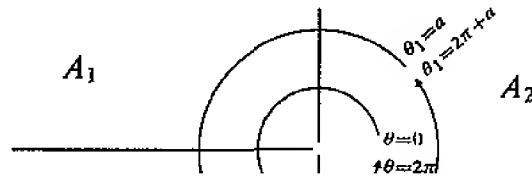
defined in Chapter 2, is C^∞ ; in fact,

$$(r, \theta): \mathbb{R}^2 - L \rightarrow \{r: r > 0\} \times (0, 2\pi)$$

is the inverse of the map

$$(a, b) \mapsto (a \cos b, a \sin b),$$

whose derivative at (a, b) has determinant equal to $a \neq 0$. By deleting a different ray L_1 we can define a different function θ_1 . Then $\theta_1 = \theta$ in the region A_1 and $\theta_1 = \theta + 2\pi$ in the region A_2 . Consequently $d\theta$ and $d\theta_1$ agree on their common



domain, so that together they define a 1-form ω on $\mathbb{R}^2 - \{0\}$. A computation (Problem 20) shows that

$$\omega = \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy.$$

The 1-form ω is usually denoted by $d\theta$, but this is an abuse of notation, since $\omega = d\theta$ only on $\mathbb{R}^2 - L$. In fact, ω is not df for any C^1 function $f: \mathbb{R}^2 - \{0\} \rightarrow \mathbb{R}$. Indeed, if $\omega = df$, then

$$df = d\theta \quad \text{on} \quad \mathbb{R}^2 - L,$$

so $d(f - \theta) = 0$ on $\mathbb{R}^2 - L$, which implies that $\partial f / \partial x = \partial \theta / \partial x$ and $\partial f / \partial y = \partial \theta / \partial y$ and hence $f = \theta + \text{constant}$ on $\mathbb{R}^2 - L$, which is impossible. Nevertheless, $d\omega = 0$ [the two relations

$$d(d\theta) = 0 \quad \text{on} \quad \mathbb{R}^2 - L$$

$$d(d\theta_1) = 0 \quad \text{on} \quad \mathbb{R}^2 - L_1$$

clearly imply that this is so]. So ω is closed, but not exact. (It is still exact in a neighborhood of any point of $\mathbb{R}^2 - \{0\}$.)

Clearly ω is also not exact in any small region containing 0. This example shows that it is the shape of the region, rather than its size, that determines whether or not a closed form is necessarily exact.

A manifold M is called (smoothly) contractible to a point $p_0 \in M$ if there is a C^∞ function

$$H: M \times [0, 1] \rightarrow M$$

such that

$$\begin{aligned} H(p, 1) &= p \\ H(p, 0) &= p_0 \end{aligned} \quad \text{for } p \in M.$$

For example, \mathbb{R}^n is smoothly contractible to $0 \in \mathbb{R}^n$; we can define

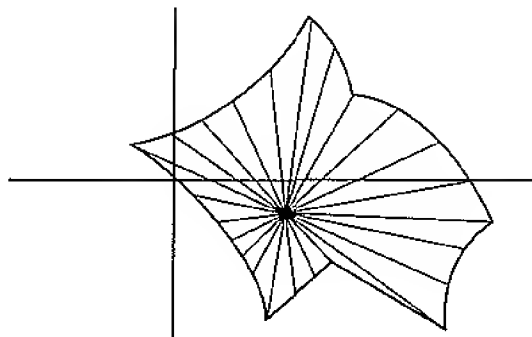
$$H: \mathbb{R}^n \times [0, 1] \rightarrow \mathbb{R}^n$$

by

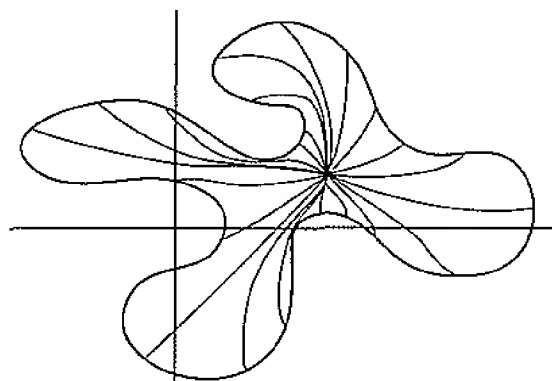
$$H(p, t) = tp.$$

More generally, $U \subset \mathbb{R}^n$ is contractible to $p_0 \in U$ if U has the property that

$p \in U$ implies $p_0 + t(p - p_0) \in U$ for $0 \leq t \leq 1$ (such a region U is called **star-shaped** with respect to p_0).



Of course, many other regions are also contractible to a point. If we think



of $[0, 1]$ as representing time, then for each time t we have a map $p \mapsto H(p, t)$ of M into itself; at time 1 this is just the identity map, and at time 0 it is the constant map.

We will show that if M is smoothly contractible to a point, then every closed form on M is exact. (By the way, this result and our investigation of the form $d\theta$ prove the intuitively obvious fact that $\mathbb{R}^2 - \{0\}$ is *not* contractible to a point; the same result holds for $\mathbb{R}^n - \{0\}$, but we will not be in a position to prove this until the next chapter.) The trick in proving our result is to analyze $M \times [0, 1]$ (for any manifold M), and pay hardly any attention at all to H .

For $t \in [0, 1]$ we define

$$i_t: M \rightarrow M \times [0, 1]$$

by

$$i_t(p) = (p, t).$$

We claim that if ω is a form on $M \times [0, 1]$ with $d\omega = 0$, then

$$i_1^* \omega - i_0^* \omega \quad \text{is exact;}$$

we will see later (and you may try to convince yourself right now) that the theorem follows trivially from this.

Consider first a 1-form ω on $M \times [0, 1]$. We will begin by working in a coordinate system on $M \times [0, 1]$. There is an obvious function t on $M \times [0, 1]$ (namely, the projection π on the second coordinate), and if (x, U) is a coordinate system on M , while π_M is the projection on M , then

$$(x^1 \circ \pi_M, \dots, x^n \circ \pi_M, t)$$

is a coordinate system on $U \times [0, 1]$. We will denote $x^i \circ \pi_M$ by \bar{x}^i , for convenience. It is easy to check (or should be) that

$$i_\alpha^* \left(\sum_{i=1}^n \omega_i d\bar{x}^i + f dt \right) = \sum_{i=1}^n \omega_i(\cdot, \alpha) dx^i,$$

where

$$\omega_i(\cdot, \alpha) \text{ denotes the function } p \mapsto \omega_i(p, \alpha).$$

Now for $\omega = \sum_{i=1}^n \omega_i d\bar{x}^i + f dt$ we have

$$d\omega = [\text{terms not involving } dt] - \sum_{i=1}^n \frac{\partial \omega_i}{\partial t} d\bar{x}^i \wedge dt + \sum_{i=1}^n \frac{\partial f}{\partial \bar{x}^i} d\bar{x}^i \wedge dt.$$

So $d\omega = 0$ implies that

$$\frac{\partial \omega_i}{\partial t} = \frac{\partial f}{\partial \bar{x}^i}.$$

Consequently,

$$\begin{aligned} \omega_i(p, 1) - \omega_i(p, 0) &= \int_0^1 \frac{\partial \omega_i}{\partial t}(p, t) dt \\ &= \int_0^1 \frac{\partial f}{\partial \bar{x}^i}(p, t) dt, \end{aligned}$$

so

$$(1) \quad \sum_{i=1}^n \omega_i(p, 1) dx^i - \sum_{i=1}^n \omega_i(p, 0) dx^i = \sum_{i=1}^n \left(\int_0^1 \frac{\partial f}{\partial \bar{x}^i}(p, t) dt \right) dx^i.$$

If we define $g: M \rightarrow \mathbb{R}$ by

$$g(p) = \int_0^1 f(p, t) dt,$$

then

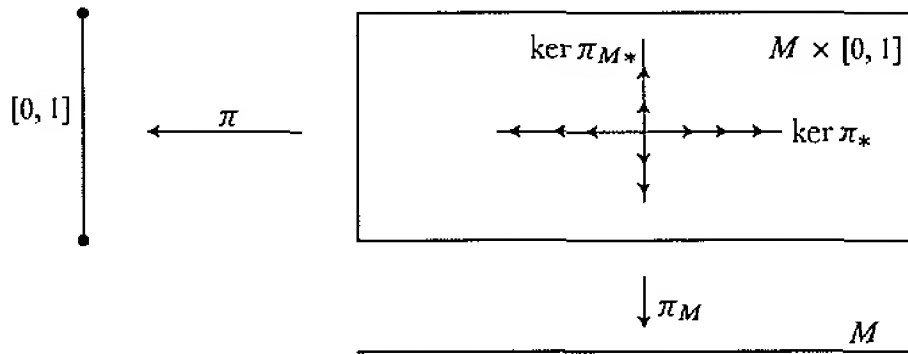
$$(2) \quad \frac{\partial g}{\partial x^i}(p) = \int_0^1 \frac{\partial f}{\partial \bar{x}^i}(p, t) dt.$$

Equations (1) and (2) show that

$$i_1^* \omega - i_0^* \omega = dg.$$

Now although we seem to be using a coordinate system, the function f , and hence g also, is really independent of the coordinate system. Notice that for the tangent space of $M \times [0, 1]$ we have

$$(*) \quad (M \times [0, 1])_{(p, t)} = \ker \pi_* \oplus \ker \pi_{M*}.$$



If a vector space V is a direct sum $V = V_1 \oplus V_2$ of two subspaces, then any $\omega \in \Omega^1(V)$ can be written

$$\omega = \omega_1 + \omega_2$$

where

$$\omega_1(v_1 + v_2) = \omega(v_1)$$

$$\omega_2(v_1 + v_2) = \omega(v_2).$$

Applying this to the decomposition (*), we write the 1-form ω on $M \times [0, 1]$ as $\omega_1 + \omega_2$; there is then a unique f with $\omega_2 = f dt$.

In general, for a k -form ω , it is easy to see (Problem 22) that we can write ω uniquely as

$$\omega = \omega_1 + (dt \wedge \eta)$$

where $\omega_1(v_1, \dots, v_k) = 0$ if some $v_i \in \ker \pi_{M*}$, and η is a $(k-1)$ -form with the

analogous property. Define a $(k-1)$ -form $I\omega$ on M as follows:

$$I\omega(p)(v_1, \dots, v_{k-1}) = \int_0^1 \eta(p, t)(i_{t*}v_1, \dots, i_{t*}v_{k-1}) dt.$$

We claim that $d\omega = 0$ implies that $i_1^*\omega - i_0^*\omega = d(I\omega)$. Actually, it is easier to find a formula for $i_1^*\omega - i_0^*\omega$ that holds even when $d\omega \neq 0$.

17. THEOREM. For any k -form ω on $M \times [0, 1]$ we have

$$i_1^*\omega - i_0^*\omega = d(I\omega) + I(d\omega).$$

(Consequently, $i_1^*\omega - i_0^*\omega = d(I\omega)$ if $d\omega = 0$.)

PROOF. Since $I\omega$ is already invariantly defined, we can just as well work in a coordinate system $(\bar{x}^1, \dots, \bar{x}^n, t)$. The operator I is clearly linear, so we just have to consider two cases.

(1) $\omega = f d\bar{x}^{i_1} \wedge \dots \wedge d\bar{x}^{i_k} = f d\bar{x}^I$. Then

$$d\omega = \frac{\partial f}{\partial t} dt \wedge d\bar{x}^I;$$

it is easy to see that

$$\begin{aligned} I(d\omega)(p) &= \left(\int_0^1 \frac{\partial f}{\partial t}(p, t) dt \right) dx^I(p) \\ &= [f(p, 1) - f(p, 0)] dx^I(p) \\ &= i_1^*\omega(p) - i_0^*\omega(p). \end{aligned}$$

Since $I\omega = 0$, this proves the result in this case.

(2) $\omega = f dt \wedge d\bar{x}^{i_1} \wedge \dots \wedge d\bar{x}^{i_{k-1}} = f dt \wedge d\bar{x}^I$. Then $i_1^*\omega = i_0^*\omega = 0$. Now

$$\begin{aligned} I(d\omega)(p) &= I\left(-\sum_{\alpha=1}^n \frac{\partial f}{\partial \bar{x}^\alpha} dt \wedge d\bar{x}^\alpha \wedge d\bar{x}^I\right)(p) \\ &= -\sum_{\alpha=1}^n \left(\int_0^1 \frac{\partial f}{\partial \bar{x}^\alpha}(p, t) dt \right) dx^\alpha \wedge dx^I \end{aligned}$$

and

$$\begin{aligned} d(I\omega) &= d\left(\int_0^1 f(p, t) dt\right) dx^I \\ &= \sum_{\alpha=1}^n \frac{\partial}{\partial x^\alpha} \left(\int_0^1 f(p, t) dt \right) dx^\alpha \wedge dx^I. \end{aligned}$$

Clearly $I(d\omega) + d(I\omega) = 0$. ♦

18. COROLLARY. If M is smoothly contractible to a point $p_0 \in M$, then every closed form ω on M is exact.

PROOF. We are given $H: M \times [0, 1] \rightarrow M$ with

$$\begin{aligned} H(p, 1) &= p \\ H(p, 0) &= p_0 \end{aligned} \quad \text{for all } p \in M.$$

Thus

$$\begin{aligned} H \circ i_1: M &\rightarrow M \quad \text{is the identity} \\ H \circ i_0: M &\rightarrow M \quad \text{is the constant map } p_0. \end{aligned}$$

So

$$\begin{aligned} \omega &= (H \circ i_1)^*(\omega) = i_1^*(H^*\omega) \\ 0 &= (H \circ i_0)^*(\omega) = i_0^*(H^*\omega). \end{aligned}$$

But

$$d(H^*\omega) = H^*(d\omega) = 0,$$

so

$$\begin{aligned} \omega - 0 &= i_1^*(H^*\omega) - i_0^*(H^*\omega) \\ &= d(I(H^*\omega)) \quad \text{by the Theorem. } \blacklozenge \end{aligned}$$

Corollary 18 is called the Poincaré Lemma by most geometers, while $d^2 = 0$ is called the Poincaré Lemma by some (I don't even know whether Poincaré had anything to do with it.) In the case of a star-shaped open subset U of \mathbb{R}^n , where we have an explicit formula for H , we can find (Problem 23) an explicit formula for $I(H^*\omega)$, for every form ω on U . Since the new form is given by an integral, we can solve the system of partial differential equations $\omega = d\eta$ explicitly in terms of integrals. There are classical theorems about vector fields in \mathbb{R}^3 which can be derived from the Poincaré Lemma and its converse (Problem 27), and originally d was introduced in order to obtain a uniform generalization of all these results. Even though the Poincaré Lemma and its converse fit very nicely into our pattern for basic theorems about differential geometry, it has always been something of a mystery to me just why d turns out to be so important. An answer to this question is provided by a theorem of Palais, *Natural Operations on Differential Forms*, Trans. Amer. Math. Soc. 92 (1959), 125–141. Suppose we have any operator D from k -forms to l -forms, such that the following diagram

commutes for every C^∞ map $f: M \rightarrow N$ [it actually suffices to assume that the diagram commutes only for diffeomorphisms f].

$$\begin{array}{ccc}
 k\text{-forms on } M & \xleftarrow{f^*} & k\text{-forms on } N \\
 D \downarrow & & \downarrow D \\
 l\text{-forms on } M & \xleftarrow{f^*} & l\text{-forms on } M
 \end{array}$$

Palais' theorem says that, with few exceptions, $D = 0$. Roughly, these exceptional cases are the following. If $k = l$, then D can be a multiple of the identity map, but nothing else. If $l = k + 1$, then D can only be some multiple of d . (As a corollary, $d^2 = 0$, since d^2 makes the above diagram commute!) There is only one other case where a non-zero D exists—when k is the dimension of M and $l = 0$. In this case, D can be a multiple of “integration”, which we discuss in the next chapter.

PROBLEMS

1. Show that if we define

$$\sigma \bullet (v_1, \dots, v_k) = (v_{\sigma^{-1}(1)}, \dots, v_{\sigma^{-1}(k)}),$$

then

$$\sigma \bullet \rho \bullet (v_1, \dots, v_k) = \sigma\rho \bullet (v_1, \dots, v_k).$$

2. Let $\overline{\text{Alt}}$ be Alt without the factor $1/k!$, and define $\omega \pi \eta = \overline{\text{Alt}}(\omega \otimes \eta)$. Show that π is *not* associative. (Try $\omega, \eta \in \Omega^1(V)$ and $\theta \in \Omega^2(V)$.)

3. Let $S' \subset S_{k+l}$ be the subgroup of all σ which leave both sets $\{1, \dots, k\}$ and $\{k+1, \dots, k+l\}$ invariant. A *cross section* of S' is a subset $K \subset S_{k+l}$ containing exactly one element from each left coset of S' .

(a) Show that for any cross section K we have

$$\omega \wedge \eta(v_1, \dots, v_{k+l}) = \sum_{\sigma \in K} \text{sgn } \sigma \cdot \omega \otimes \eta(v_{\sigma(1)}, \dots, v_{\sigma(k+l)}).$$

This definition may be used even in a field of finite characteristic.

(b) Show from this definition that $\omega \wedge \eta$ is alternating, and $\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega$. (Proving associativity is quite messy.)

(c) A permutation $\sigma \in S_{k+l}$ is called a *shuffle permutation* if $\sigma(1) < \sigma(2) < \dots < \sigma(k)$ and $\sigma(k+1) < \sigma(k+2) < \dots < \sigma(k+l)$. Show that the set of all shuffle permutations is a cross section of S' .

4. For $v \in V$ and $\omega \in \Omega^k(V)$, we define the contraction $v \lrcorner \omega \in \Omega^{k-1}(V)$ by

$$(v \lrcorner \omega)(v_1, \dots, v_{k-1}) = \omega(v, v_1, \dots, v_{k-1}).$$

This is sometime also called the *inner product* and the notation $i_v \omega$ is also used.

(a) Show that

$$v \lrcorner (w \lrcorner \omega) = -w \lrcorner (v \lrcorner \omega).$$

(b) Show that if v_1, \dots, v_n is a basis of V with dual basis ϕ_1, \dots, ϕ_n , then

$$v_j \lrcorner (\phi_{i_1} \wedge \dots \wedge \phi_{i_k}) = \begin{cases} 0 & j \neq \text{any } i_\alpha \\ (-1)^{\alpha-1} \phi_{i_1} \wedge \dots \wedge \widehat{\phi_{i_\alpha}} \wedge \dots \wedge \phi_{i_k} & \text{if } j = i_\alpha. \end{cases}$$

(c) Show that for $\omega_1 \in \Omega^k(V)$ and $\omega_2 \in \Omega^l(V)$ we have

$$v \lrcorner (\omega_1 \wedge \omega_2) = (v \lrcorner \omega_1) \wedge \omega_2 + (-1)^k \omega_1 \wedge (v \lrcorner \omega_2).$$

(Use (b) and linearity of everything.)

(d) Formula (c) can be used to give a definition of $\omega_1 \wedge \omega_2$ by induction on $k + l$ (which works for vector spaces over any field): If \wedge is defined for forms of degree adding up to $< k + l$, we define

$$\begin{aligned}\omega_1 \wedge \omega_2(v_1, \dots, v_{k+l}) &= [(v_1 \lrcorner \omega_1) \wedge \omega_2](v_2, \dots, v_{k+l}) \\ &\quad + (-1)^k [\omega_1 \wedge (v_1 \lrcorner \omega_2)](v_2, \dots, v_{k+l}).\end{aligned}$$

Show that with this definition $\omega_1 \wedge \omega_2$ is skew-symmetric (it is only necessary to check that interchanging v_1 and v_2 changes the sign of the right side).

(e) Prove by induction that \wedge is bilinear and that $\omega_1 \wedge \omega_2 = (-1)^{kl} \omega_2 \wedge \omega_1$.

(f) If X is a vector field on M and ω a k -form on M we define a $(k-1)$ -form $X \lrcorner \omega$ by

$$(X \lrcorner \omega)(p) = X(p) \lrcorner \omega(p).$$

Show that if ω_1 is a k -form, then

$$X \lrcorner (\omega_1 \wedge \omega_2) = (X \lrcorner \omega_1) \wedge \omega_2 + (-1)^k \omega_1 \wedge (X \lrcorner \omega_2).$$

5. Show that n functions $f_1, \dots, f_n: M \rightarrow \mathbb{R}$ form a coordinate system in a neighborhood of $p \in M$ if and only if $df_1 \wedge \dots \wedge df_n(p) \neq 0$.

6. An element $\omega \in \Omega^k(V)$ is called **decomposable** if $\omega = \phi_1 \wedge \dots \wedge \phi_k$ for some $\phi_i \in V^* = \Omega^1(V)$.

(a) If $\dim V \leq 3$, then every $\omega \in \Omega^2(V)$ is decomposable.

(b) If $\phi_i, i = 1, \dots, 4$ are independent, then $\omega = (\phi_1 \wedge \phi_2) + (\phi_3 \wedge \phi_4)$ is not decomposable. *Hint*: Look at $\omega \wedge \omega$.

7. For any $\omega \in \Omega^k(V)$, we define the **annihilator** of ω to be

$$\text{Ann}(\omega) = \{\phi \in V^*: \phi \wedge \omega = 0\}.$$

(a) Show that

$$\dim \text{Ann}(\omega) \leq k,$$

and that equality holds if and only if ω is decomposable.

(b) Every subspace of V^* is $\text{Ann}(\omega)$ for some decomposable ω , which is unique up to a multiplicative constant.

(c) If ω_1 and ω_2 are decomposable, then $\text{Ann}(\omega_1) \subset \text{Ann}(\omega_2)$ if and only if $\omega_2 = \omega_1 \wedge \eta$ for some η .

(d) If ω_i are decomposable, then $\text{Ann}(\omega_1) \cap \text{Ann}(\omega_2) = \{0\}$ if and only if $\omega_1 \wedge \omega_2 \neq 0$. In this case,

$$\text{Ann}(\omega_1) + \text{Ann}(\omega_2) = \text{Ann}(\omega_1 \wedge \omega_2).$$

(e) If V has dimension n , then any $\omega \in \Omega^{n-1}(V)$ is decomposable.

(f) Since $v_i \in V$ can be regarded as elements of V^{**} , we can consider $v_1 \wedge \dots \wedge v_k \in \Omega^k(V^*)$. Reformulate parts (a)–(d) in terms of this \wedge product.

8. (a) Let $\omega \in \Omega^2(V)$. Show that there is a basis ϕ_1, \dots, ϕ_n of V^* such that

$$\omega = (\phi_1 \wedge \phi_2) + \dots + (\phi_{2r-1} \wedge \phi_{2r}).$$

Hint: If

$$\omega = \sum_{i < j} a_{ij} \psi_i \wedge \psi_j,$$

choose ϕ_1 involving $\psi_1, \psi_3, \dots, \psi_n$ and ϕ_2 involving ψ_2, \dots, ψ_n so that

$$\omega = \phi_1 \wedge \phi_2 + \omega',$$

where ω' does not involve ψ_1 or ψ_2 .

(b) Show that the r -fold wedge product $\omega \wedge \dots \wedge \omega$ is non-zero and decomposable, and that the $(r+1)$ -fold wedge product is 0. Thus r is well-determined; it is called the **rank** of ω .

(c) If $\omega = \sum_{i < j} a_{ij} \psi_i \wedge \psi_j$, show that the rank of ω is the rank of the matrix (a_{ij}) .

9. If v_1, \dots, v_n is a basis for V and $w_i = \sum_{j=1}^n \alpha_{ji} v_j$, show that

$$\det(\alpha_{ij}) w_1^* \wedge \dots \wedge w_n^* = v_1^* \wedge \dots \wedge v_n^*.$$

10. Let $A = (a_{ij})$ be an $n \times n$ matrix. Let $1 \leq p \leq n$ be fixed, and let $q = n - p$. For $H = h_1 < \dots < h_p$ and $K = k_1 < \dots < k_q$, let

$$B^H = \det \begin{pmatrix} a_{1,h_1} & \dots & a_{1,h_p} \\ \vdots & & \vdots \\ a_{p,h_1} & \dots & a_{p,h_p} \end{pmatrix}, \quad C^K = \det \begin{pmatrix} a_{p+1,k_1} & \dots & a_{p+1,k_q} \\ \vdots & & \vdots \\ a_{n,k_1} & \dots & a_{n,k_q} \end{pmatrix}.$$

(a) If v_1, \dots, v_n is a basis of V and

$$w_i = \sum_{j=1}^n a_{ji} v_j,$$

show that

$$\begin{aligned} w_1 \wedge \dots \wedge w_p &= \sum_H B^H v_H \\ w_{p+1} \wedge \dots \wedge w_n &= \sum_K C^K v_K. \end{aligned}$$

(b) Let $H' = \{1, \dots, n\} - H$ (arranged in increasing order). Show that

$$v_H \wedge v_K = \begin{cases} 0 & K \neq H' \\ e_{H, H'} v_1 \wedge \dots \wedge v_n & K = H', \end{cases}$$

where $e_{H, H'}$ is the sign of the permutation

$$\begin{pmatrix} 1 & \dots & \dots & \dots & \dots & \dots & n \\ h_1, h_2, \dots, h_p, k_1, \dots, k_q \end{pmatrix}.$$

(c) Prove “Laplace’s expansion”

$$\det A = \sum_H e_{H, H'} B^H C^{H'}.$$

11. (Cartan’s Lemma) Let $\phi_1, \dots, \phi_k \in V^*$ be independent and suppose that $\psi_1, \dots, \psi_k \in V^*$ satisfy

$$(\phi_1 \wedge \psi_1) + \dots + (\phi_k \wedge \psi_k) = 0.$$

Then

$$\psi_i = \sum_{j=1}^k a_{ji} \phi_j, \quad \text{where } a_{ji} = a_{ij}.$$

12. In addition to forms, we can consider sections of bundles constructed from TM using Ω and other operations. For example, if $\xi = \pi: E \rightarrow B$ is a vector bundle, we can consider $\Omega^k(\xi^*)$, the bundle whose fibre at p is $\Omega^k([\pi^{-1}(p)]^*)$. Since we can regard

$$\frac{\partial}{\partial x^i} \quad \text{as an element of } (M_p)^{**},$$

any section of $\Omega^n(T^*M)$ can be written locally as

$$h \frac{\partial}{\partial x^1} \wedge \dots \wedge \frac{\partial}{\partial x^n}.$$

(a) Show that if

$$g \frac{\partial}{\partial y^1} \wedge \dots \wedge \frac{\partial}{\partial y^n} = h \frac{\partial}{\partial x^1} \wedge \dots \wedge \frac{\partial}{\partial x^n},$$

then

$$h = g \cdot \left[\det \left(\frac{\partial y^i}{\partial x^j} \right) \right]^{-1}.$$

This shows that sections of $\Omega^n(T^*M)$ are the geometric objects corresponding to the (even) relative scalars of weight -1 in Problem 4-10.

(b) Let $\mathcal{T}_l^{k[m]}(V)$ denote the vector space of all multilinear functions

$$\underbrace{V \times \cdots \times V}_k \times \underbrace{V^* \times \cdots \times V^*}_l \rightarrow \Omega^m(V).$$

Show that sections of $\mathcal{T}_l^{k[n]}(TM)$ correspond to (even) relative tensors of type $\binom{k}{l}$ and weight 1. (Notice that if v_1, \dots, v_n is a basis for V , then elements of $\Omega^n(V)$ can be represented by real numbers [times the element $v^*_1 \wedge \cdots \wedge v^*_n$].)

(c) If $\mathcal{T}_{l[m]}^k(V)$ is defined similarly, except that $\Omega^m(V)$ is replaced by $\Omega^m(V^*)$, show that sections of $\mathcal{T}_{l[m]}^k(TM)$ correspond to (even) relative tensors of type $\binom{k}{l}$ and weight -1 .

(d) Show that the covariant relative tensor of type $\binom{0}{n}$ and weight 1 defined in Problem 4-10, with components $\varepsilon^{i_1 \dots i_n}$, corresponds to the map

$$\underbrace{V^* \times \cdots \times V^*}_n \rightarrow \Omega^n(V)$$

given by $(\phi_1, \dots, \phi_n) \mapsto \phi_1 \wedge \cdots \wedge \phi_n$. Interpret the relative tensor with components $\varepsilon_{i_1 \dots i_n}$ similarly.

(e) Suppose $\Omega^{n;w}(V)$ denotes all functions $\eta: V \times \cdots \times V \rightarrow \mathbb{R}$ which are of the form

$$\eta(v_1, \dots, v_n) = [\omega(v_1, \dots, v_n)]^w \quad w \text{ an integer}$$

for some $\omega \in \Omega^n(V)$. Let $\mathcal{T}_l^{k[n;w]}(V)$ be defined like $\mathcal{T}_l^{k[n]}$, except that $\Omega^n(V)$ is replaced by $\Omega^{n;w}(V)$. Show that sections of $\mathcal{T}_l^{k[n;w]}(TM)$ correspond to (even) relative tensors of type $\binom{k}{l}$ and weight w . Similarly for $\mathcal{T}_{l[n;w]}^k$.

(f) For those who know about tensor products $V \otimes W$ and exterior algebras $\Lambda^k(V)$, these results can all be restated. We can identify $\mathcal{T}_l^k(V)$ with

$$\bigotimes^k V^* \otimes \bigotimes^l V = \underbrace{V^* \otimes \cdots \otimes V^*}_k \otimes \underbrace{V \otimes \cdots \otimes V}_l.$$

Since $\Omega^m(V) \approx \Lambda^m(V^*) \approx [\Lambda^m(V)]^*$, we can identify

$$\begin{aligned} \mathcal{T}_l^{k[m]}(V) & \text{ with } \bigotimes^k V^* \otimes \bigotimes^l V \otimes \Lambda^m(V) \\ \mathcal{T}_{l[m]}^k(V) & \text{ with } \bigotimes^k V^* \otimes \bigotimes^l V \otimes \Lambda^m(V^*). \end{aligned}$$

Consider, more generally,

$$\begin{aligned}\mathcal{T}_l^{k[m;w]}(V) &= \bigotimes^k V^* \otimes \bigotimes^l V \otimes \bigotimes^w \Lambda^m(V) \\ \mathcal{T}_{l[m;w]}^k(V) &= \bigotimes^k V^* \otimes \bigotimes^l V \otimes \bigotimes^w \Lambda^m(V^*).\end{aligned}$$

Noting that $\Lambda^n(V) \otimes \cdots \otimes \Lambda^n(V)$ is always 1-dimensional, show that sections of $\mathcal{T}_l^{k[n;w]}(TM)$ and $\mathcal{T}_{l[n;w]}^k(TM)$ correspond to (even) relative tensors of type $\binom{k}{l}$ and weight w and $-w$, respectively.

13. (a) If V has dimension n and $A: V \rightarrow V$ is a linear transformation, then the map $A^*: \Omega^n(V) \rightarrow \Omega^n(V)$ must be multiplication by some constant c . Show that $c = \det A$. (This may be used as a definition of $\det A$.)

(b) Conclude that $\det AB = (\det A)(\det B)$.

14. Recall that the characteristic polynomial of $A: V \rightarrow V$ is

$$\begin{aligned}\chi(\lambda) &= \det(\lambda I - A) \\ &= \lambda^n - (\text{trace } A)\lambda^{n-1} + \cdots + (-1)^n \det A \\ &= \lambda^n - c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + (-1)^n c_n.\end{aligned}$$

(a) Show that $c_k = \text{trace of } A^*: \Omega^k(V) \rightarrow \Omega^k(V)$.

(b) Conclude that $c_k(AB) = c_k(BA)$.

(c) Let $\delta_{i_1 \dots i_k}^{j_1 \dots j_k}$ be as defined in Problem 4-5(xiii). If $A: V \rightarrow V$ has a matrix (a_i^j) (with respect to some basis), show that

$$c_k(A) = \frac{1}{k!} \sum_{\substack{i_1, \dots, i_k \\ j_1, \dots, j_k}} a_{i_1}^{j_1} a_{i_2}^{j_2} \cdots a_{i_k}^{j_k} \delta_{j_1 \dots j_k}^{i_1 \dots i_k}.$$

Thus, if δ is as defined on page 130, and A is a tensor of type $\binom{1}{1}$, then the function $p \mapsto c_k(A(p))$ can be defined as a $(2k)$ -fold contraction of

$$\underbrace{A \otimes \cdots \otimes A}_{k \text{ times}} \otimes \delta.$$

15. Let $P(X_{ij})$ be a polynomial in n^2 variables. For every $n \times n$ matrix $A = (a_{ij})$ we then have a number $P(a_{ij})$. Call P invariant if $P(A) = P(BAB^{-1})$ for all A and all invertible B . This problem outlines a proof that any invariant P is a polynomial in the polynomials c_1, \dots, c_n defined in Problem 14. We will

need the algebraic result that any symmetric polynomial $Q(y_1, \dots, y_n)$ in the n variables y_1, \dots, y_n can be written as a polynomial in $\sigma_1, \dots, \sigma_n$, where σ_i is the i^{th} elementary symmetric polynomial of y_1, \dots, y_n . Recall that the σ_i can be defined by the equation

$$\prod_{i=1}^n (y - y_i) = y^n - \sigma_1 y^{n-1} + \dots + (-1)^n \sigma_n.$$

Thus, they are the coefficients, up to sign, of the polynomial with roots y_1, \dots, y_n . Since the eigenvalues $\lambda_1, \dots, \lambda_n$ of a matrix A are, by definition, the roots of the polynomial $\chi(\lambda)$, it follows that

$$c_i(A) = \sigma_i(\lambda_1, \dots, \lambda_n).$$

We will first consider matrices A over the complex numbers \mathbb{C} (the coefficients of P may also be complex).

(a) Define $Q(y_1, \dots, y_n)$ to be $P(A)$ where A is the diagonal matrix

$$\begin{pmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{pmatrix}.$$

Then there is a polynomial R such that

$$Q(y_1, \dots, y_n) = R(\sigma_1(y_1, \dots, y_n), \dots, \sigma_n(y_1, \dots, y_n)).$$

The polynomial R has real coefficients if P does.

(b) $P(A) = R(c_1(A), \dots, c_n(A))$ for all diagonalizable A .

(c) The discriminant $D(A)$ is defined as $\prod_{i \neq j} (\lambda_i - \lambda_j)^2$, where λ_i are the eigenvalues of A . Show that $D(A)$ can be written as a polynomial in the entries of A .

(d) Show that $P(A) = R(c_1(A), \dots, c_n(A))$ whenever $D(A) \neq 0$. Conclude, by continuity, that the equation holds for all matrices A over \mathbb{C} . (This last conclusion follows even if \mathbb{C} is replaced by some other field, since the set where $D \neq 0$ is Zariski-dense; this is “the principle of irrelevance of algebraic inequalities”, compare pg. V.375.)

Now suppose that the coefficients of P are real and that $P(A) = P(BAB^{-1})$ for all real A and real invertible B .

(e) The same equation holds for complex A and complex invertible B . (Regard the equation as n^2 polynomial equations in the a_{ij} and b_{ij} .)

16. (a) Let v_1, \dots, v_n be a basis for V , and let $w_1, \dots, w_k \in V$ be given by

$$w_i = \sum_{j=1}^n \alpha_{ji} v_j.$$

For $\omega \in \Omega^k(V)$ show that

$$\omega(w_1, \dots, w_k) = \sum_{I=i_1 < \dots < i_k} \alpha_I \omega(v_{i_1}, \dots, v_{i_k}),$$

where α_I is the determinant of the $k \times k$ submatrix of (α_{ij}) obtained by selecting rows i_1, \dots, i_k .

(b) Generalize Theorem 7 and Corollary 8 to k -forms.

(c) Check directly from (b) that the definition of d does not depend on the coordinate system.

17. Show that $d(\sum_{i < j} \alpha_{ij} dx^i \wedge dx^j) = 0$ if and only if

$$\frac{\partial \alpha_{ij}}{\partial x^k} - \frac{\partial \alpha_{ik}}{\partial x^j} + \frac{\partial \alpha_{jk}}{\partial x^i} = 0 \quad \text{for all } i < j < k.$$

18. In Problem 5-14 we defined $L_X A$ for any tensor field A .

(a) Show that if ω is a k -form, then so is $L_X \omega$.

(b) Show that

$$L_X(\omega_1 \wedge \omega_2) = L_X \omega_1 \wedge \omega_2 + \omega_1 \wedge L_X \omega_2.$$

(c) Using 5-14(e), show that

$$\begin{aligned} X(\omega(X_1, \dots, X_k)) &= L_X(\omega(X_1, \dots, X_k)) \\ &= L_X \omega(X_1, \dots, X_k) \\ &\quad + \sum_{i=1}^k (-1)^{i+1} \omega([X, X_i], X_1, \dots, \widehat{X_i}, \dots, X_k). \end{aligned}$$

(d) Deduce the following two expressions:

$$\begin{aligned} d\omega(X_1, \dots, X_{k+1}) &= \sum_{i=1}^{k+1} (-1)^{i+1} L_{X_i} \omega(X_1, \dots, \widehat{X_i}, \dots, X_{k+1}) \\ &\quad + \sum_{i < j} (-1)^{i+j+1} \omega([X_i, X_j], X_1, \dots, \widehat{X_i}, \dots, \widehat{X_j}, \dots, X_{k+1}) \end{aligned}$$

$$\begin{aligned}
d\omega(X_1, \dots, X_{k+1}) \\
&= \frac{1}{2} \sum_{i=1}^{k+1} (-1)^{i+1} \{X_i(\omega(X_1, \dots, \widehat{X}_i, \dots, X_{k+1})) \\
&\quad + L_{X_i} \omega(X_1, \dots, \widehat{X}_i, \dots, X_{k+1})\}
\end{aligned}$$

(e) Show that

$$X \lrcorner d\omega = L_X \omega - d(X \lrcorner \omega),$$

i.e.,

$$d\omega(X_1, \dots, X_{k+1}) = (L_{X_1} \omega)(X_2, \dots, X_{k+1}) - d(X_1 \lrcorner \omega)(X_2, \dots, X_{k+1}).$$

(This may be used to give an inductive definition of d .)

(f) Using (e), show that $d(L_X \omega) = L_X(d\omega)$.

19. Let a_{ij} be n^2 functions on \mathbb{R}^n with $a_{ij} = a_{ji}$. Show that in order for there to be functions u_1, \dots, u_n in a neighborhood of any point in \mathbb{R}^n with

$$a_{ij} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x^j} + \frac{\partial u_j}{\partial x^i} \right)$$

it is necessary and sufficient that

$$\frac{\partial^2 a_{ij}}{\partial x^k \partial x^l} - \frac{\partial^2 a_{ik}}{\partial x^j \partial x^l} = \frac{\partial^2 a_{lj}}{\partial x^k \partial x^i} - \frac{\partial^2 a_{lk}}{\partial x^j \partial x^i} \quad \text{for all } i, j, k, l.$$

Hint: First set up partial differential equations for the functions $f_{jk} = \partial u_j / \partial x^k - \partial u_k / \partial x^j$, and use Theorem 6-1.

20. Compute that

$${}^{\text{“}}d\theta{}^{\text{”}} = \frac{x dy - y dx}{x^2 + y^2}.$$

(At most places $\theta = \arctan y/x$ [+ a constant].)

21. (a) If ω is a 1-form $f dx$ on $[0, 1]$ with $f(0) = f(1)$, show that there is a unique number λ such that $\omega - \lambda dx = dg$ for some function g with $g(0) = g(1)$.

Hint: Integrate the equation $\omega - \lambda dx = dg$ on $[0, 1]$ to find λ .

(b) Let $i: S^1 \rightarrow \mathbb{R}^2 - \{0\}$ be the inclusion, and let $\sigma' = i^*(d\theta)$. If $c: [0, 1] \rightarrow S^1$ is

$$c(x) = (\cos 2\pi x, \sin 2\pi x),$$

show that

$$c^*(\sigma') = 2\pi dx.$$

(c) If ω is a closed 1-form on S^1 show that there is a unique number λ such that $\omega - \lambda \sigma'$ is exact.

22. (a) Show that every $\omega \in \Omega^k(V_1 \oplus V_2)$ can be written as a sum of forms $\omega_1 \wedge \omega_2$ where ω_1 has degree α and ω_2 has degree $\beta = k - \alpha$ and

$$\omega_1(v_1, \dots, v_\alpha) = 0 \quad \text{if some } v_i \in V_2$$

$$\omega_2(v_1, \dots, v_\beta) = 0 \quad \text{if some } v_i \in V_1.$$

(b) If $\dim V_2 = 1$, and $0 \neq \lambda \in V_2^*$, then ω can be written uniquely as $\omega_1 + (\omega_2 \wedge \lambda)$, where ω_1 is a k -form and ω_2 is a $(k - 1)$ -form such that

$$\omega_1(v_1, \dots, v_k) = 0 \quad \text{if some } v_i \in V_2$$

$$\omega_2(v_1, \dots, v_{k-1}) = 0 \quad \text{if some } v_i \in V_2.$$

23. Let $U \subset \mathbb{R}^n$ be an open set star-shaped with respect to 0, and define $H: U \times [0, 1] \rightarrow U$ by $H(p, t) = tp$. If

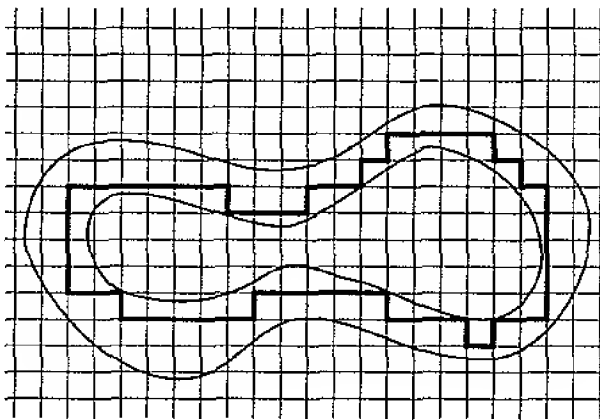
$$\omega = \sum_{i_1 < \dots < i_k} \omega_{i_1 \dots i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k}$$

on U , show that

$$I(H^*\omega)$$

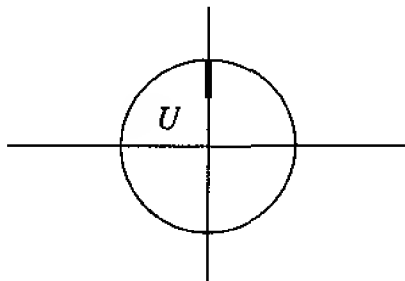
$$= \sum_{i_1 < \dots < i_k} \sum_{\alpha=1}^k (-1)^{\alpha-1} \left(\int_0^1 t^{k-1} \omega_{i_1 \dots i_k}(tx) dt \right) x^{i_\alpha} dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\alpha}} \wedge \dots \wedge dx^{i_k}.$$

24. (a) Let $U \subset \mathbb{R}^2$ be a bounded open set such that $\mathbb{R}^2 - U$ is connected. Show that U is diffeomorphic to \mathbb{R}^2 , and hence smoothly contractible to a point. (The converse is proved in Problem 8-9.) *Hint*: Obtain U as an increasing union of sets, the k^{th} set being a finite union of squares containing the set of points in U whose distance from boundary U is $\leq 1/k$.



(b) Find a bounded open set $U \subset \mathbb{R}^3$ such that $\mathbb{R}^3 - U$ is connected, but U is not contractible to a point.

25. Let $U \subset \mathbb{R}^n$ be an open set star-shaped with respect to 0. Is U homeomorphic to \mathbb{R}^n ? (It would certainly appear so, but the “obvious” proof does not work, since the length of rays from 0 to the boundary of the set could vary discontinuously.)



26. Let $\langle \cdot, \cdot \rangle$ be the usual inner product on \mathbb{R}^n ,

$$\langle a, b \rangle = \sum_{i=1}^n a^i b^i.$$

(a) If $v_1, \dots, v_{n-1} \in \mathbb{R}^n$, show that there is a unique vector $v_1 \times \dots \times v_{n-1} \in \mathbb{R}^n$ with

$$\langle v_1 \times \dots \times v_{n-1}, w \rangle = \det \begin{pmatrix} w \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix} \text{ for all } w \in \mathbb{R}^n.$$

(b) Show that $v_1 \times \dots \times v_{n-1} \in \Omega^{n-1}(\mathbb{R}^n)$, and express it in terms of the e^*_i , using the expansion of a matrix by minors.

(c) For \mathbb{R}^3 show that

$$v \times w = (v^2 w^3 - v^3 w^2, v^3 w^1 - v^1 w^3, v^1 w^2 - v^2 w^1).$$

(First find all $e_i \times e_j$.)

27. (a) If $f: \mathbb{R}^n \rightarrow \mathbb{R}$, define a vector field $\text{grad } f$, the **gradient** of f , on \mathbb{R}^n by

$$\text{grad } f = \sum_{i=1}^n \frac{\partial f}{\partial x^i} \cdot \frac{\partial}{\partial x^i} = \sum_{i=1}^n D_i f \cdot \frac{\partial}{\partial x^i}.$$

Introducing the formal symbolism

$$\nabla = \sum_{i=1}^n D_i \frac{\partial}{\partial x^i},$$

we can write $\text{grad } f = \nabla f$. If $(\text{grad } f)(p) = w_p$, show that

$$D_v f(p) = \langle v, w \rangle,$$

where $D_v f(p)$ denotes the directional derivative in the direction v at p (or simply $v_p(f)$, if we regard $v_p \in \mathbb{R}^n_p$). Conclude that $\nabla f(p)$ is the direction in which f is changing fastest at p .

(b) If $X = \sum_{i=1}^n a^i \partial/\partial x^i$ is a vector field on \mathbb{R}^n , we define the **divergence** of X as

$$\text{div } X = \sum_{i=1}^n \frac{\partial a^i}{\partial x^i}.$$

(Symbolically, we can write $\text{div } X = \langle \nabla, X \rangle$.) We also define, for $n = 3$,

$$\begin{aligned} \text{curl } X (= \nabla \times X) \\ = \left(\frac{\partial a^3}{\partial x^2} - \frac{\partial a^2}{\partial x^3} \right) \frac{\partial}{\partial x^1} + \left(\frac{\partial a^1}{\partial x^3} - \frac{\partial a^3}{\partial x^1} \right) \frac{\partial}{\partial x^2} + \left(\frac{\partial a^2}{\partial x^1} - \frac{\partial a^1}{\partial x^2} \right) \frac{\partial}{\partial x^3}. \end{aligned}$$

Define forms

$$\begin{aligned} \omega_X &= a^1 dx + a^2 dy + a^3 dz \\ \eta_X &= a^1 dy \wedge dz + a^2 dz \wedge dx + a^3 dx \wedge dy. \end{aligned}$$

Show that

$$\begin{aligned} df &= \omega_{\text{grad } f} \\ d(\omega_X) &= \eta_{\text{curl } X} \\ d(\eta_X) &= (\text{div } X) dx \wedge dy \wedge dz. \end{aligned}$$

(c) Conclude that

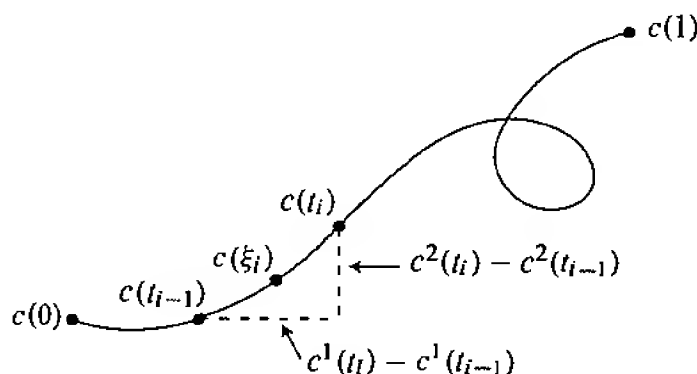
$$\begin{aligned} \text{curl grad } f &= 0 \\ \text{div curl } X &= 0. \end{aligned}$$

(d) If X is a vector field on a star-shaped open set $U \subset \mathbb{R}^n$ and $\text{curl } X = 0$, then $X = \text{grad } f$ for some function $f: U \rightarrow \mathbb{R}$. Similarly, if $\text{div } X = 0$, then $X = \text{curl } Y$ for some vector field Y on U .

CHAPTER 8

INTEGRATION

The basic concept of this chapter generalizes line and surface integrals, which first arose from very physical considerations. Suppose, for example, that $c: [0, 1] \rightarrow \mathbb{R}^2$ is a curve and $\omega = f dx + g dy$ is a 1-form on \mathbb{R}^2 (where $f, g: \mathbb{R}^2 \rightarrow \mathbb{R}$, and x and y denote the coordinate functions on \mathbb{R}^2). If we choose a partition $0 = t_0 < \cdots < t_n = 1$ of $[0, 1]$, then we can divide the curve c into n pieces, the i^{th} piece going from $c(t_{i-1})$ to $c(t_i)$. When the differences $t_i - t_{i-1}$ are small, each such piece is approximately a straight segment, with



horizontal projection $c^1(t_i) - c^1(t_{i-1})$ and vertical projection $c^2(t_i) - c^2(t_{i-1})$. We can choose points $c(\xi_i)$ on each piece by choosing points $\xi_i \in [t_{i-1}, t_i]$. For each partition P and each such choice $\xi = (\xi_1, \dots, \xi_n)$, consider the sum

$$S(P, \xi) = \sum_{i=1}^n f(c(\xi_i)) [c^1(t_i) - c^1(t_{i-1})] + g(c(\xi_i)) [c^2(t_i) - c^2(t_{i-1})].$$

If these sums approach a limit as the “mesh” $\|P\|$ of P approaches 0, that is, as the maximum of $t_i - t_{i-1}$ approaches 0, then the limit is denoted by

$$\int_c f dx + g dy.$$

(This is a complicated limit. To be precise, if $\|P\| = \max_i \{t_i - t_{i-1}\}$, then the equation

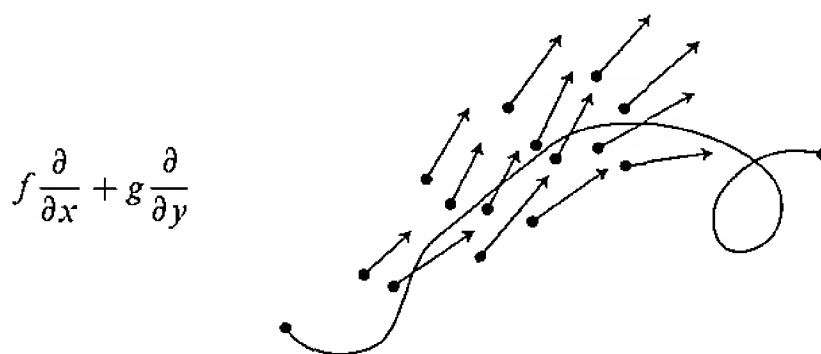
$$\lim_{\|P\| \rightarrow 0} S(P, \xi) = \int_c f dx + g dy$$

means: for all $\varepsilon > 0$, there is a $\delta > 0$ such that for all partitions P with $\|P\| < \delta$, we have

$$\left| S(P, \xi) - \int_c f dx + g dy \right| < \varepsilon$$

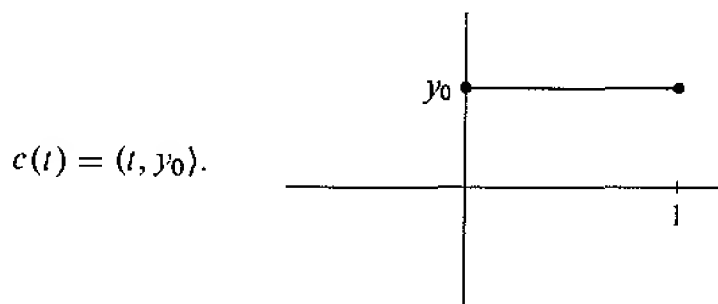
for all choices ξ for P .)

The limit which we have just defined is called a “line integral”; it has a natural physical interpretation. If we consider a “force field” on \mathbb{R}^2 , described by the vector field



then $S(P, \xi)$ is the “work” involved in moving a unit mass along the curve c in the case where c is actually a straight line between t_{i-1} and t_i and f and g are constant along these straight line segments; the limit is the natural definition of the work done in the general case. (In classical terminology, the differential $f dx + g dy$ would be described as the work done by the force field on an “infinitely small” displacement with components dx, dy ; the integral is the “sum” of these infinitely small displacements.)

Before worrying about how to compute this limit, consider the special case where



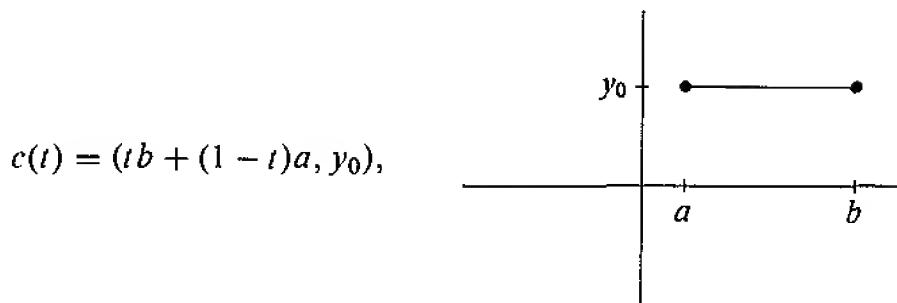
In this case, $c^1(t_i) - c^1(t_{i-1}) = t_i - t_{i-1}$, while $c^2(t_i) - c^2(t_{i-1}) = 0$, so

$$S(P, \xi) = \sum_{i=1}^n f(\xi_i, y_0)(t_i - t_{i-1}).$$

These sums approach

$$\int_c f dx + g dy = \int_0^1 f(x, y_0) dx.$$

On the other hand, if



then $c^1(t_i) - c^1(t_{i-1}) = (b - a)(t_i - t_{i-1})$, so

$$S(P, \xi) = (b - a) \cdot \sum_{i=1}^n f(\xi_i b + (1 - \xi_i)a, y_0)(t_i - t_{i-1}).$$

These sums approach

$$(b - a) \int_0^1 f(xb + (1 - x)a, y_0) dx = \int_a^b f(x, y_0) dx.$$

In general, for any curve c , we have, by the mean value theorem,

$$\begin{aligned} c^1(t_i) - c^1(t_{i-1}) &= c^{1'}(\alpha_i)(t_i - t_{i-1}) & \alpha_i &\in [t_{i-1}, t_i] \\ c^2(t_i) - c^2(t_{i-1}) &= c^{2'}(\beta_i)(t_i - t_{i-1}) & \beta_i &\in [t_{i-1}, t_i]. \end{aligned}$$

So

$$S(P, \xi) = \sum_{i=1}^n \{f(c(\xi_i))c^{1'}(\alpha_i) + g(c(\xi_i))c^{2'}(\beta_i)\} (t_i - t_{i-1}).$$

A somewhat messy argument (Problem 1) shows that these sums approach what it looks like they should approach, namely

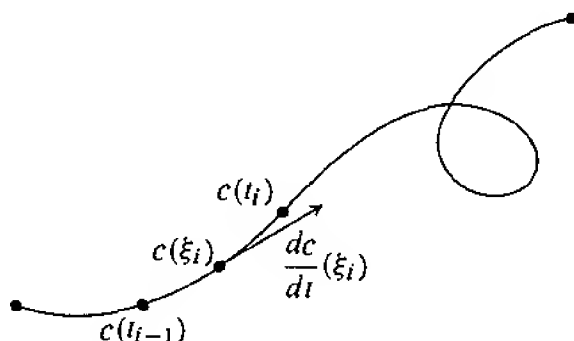
$$\int_0^1 [f(c(t))c^{1'}(t) + g(c(t))c^{2'}(t)] dt.$$

Physicists' notation (or abuse thereof) makes it easy to remember this result. The components c^1, c^2 of c are denoted simply by x and y [i.e., x denotes

$x \circ c$ and y denotes $y \circ c$; this is indicated classically by saying “let $x = x(t)$, $y = y(t)$ ”. The above integral is then written

$$\int_c f dx + g dy = \int_0^1 \left[f(x, y) \frac{dx}{dt} + g(x, y) \frac{dy}{dt} \right] dt.$$

In preference to this physical interpretation of “line integrals”, we can introduce a more geometrical interpretation. Recall that $dc/dt(\xi_i)$ denotes the



tangent vector of c at time ξ_i . Then the sums

$$\begin{aligned} (*) \quad & \sum_{i=1}^n \omega(c(\xi_i)) \left(\frac{dc}{dt}(\xi_i) \right) \cdot (t_i - t_{i-1}) \\ &= \sum_{i=1}^n [f(c(\xi_i))c^1(\xi_i) + g(c(\xi_i))c^2(\xi_i)] \cdot (t_i - t_{i-1}) \end{aligned}$$

clearly also approach

$$\int_0^1 [f(c(t))c^1(t) + g(c(t))c^2(t)] dt.$$

Consider the special case where c goes with constant velocity on each (t_{i-1}, t_i) .



If we choose any $\xi_i \in (t_{i-1}, t_i)$, then

$$\begin{aligned} \text{length of } \frac{dc}{dt}(\xi_i) &= \text{the constant speed on } (t_{i-1}, t_i) \\ &= \frac{\text{length of the segment from } c(t_{i-1}) \text{ to } c(t_i)}{t_i - t_{i-1}}, \end{aligned}$$

so

$$\left[\text{length of } \frac{dc}{dt}(\xi_i) \right] \cdot (t_i - t_{i-1}) = \text{length of segment from } c(t_{i-1}) \text{ to } c(t_i).$$

In this case,

$$\sum_{i=1}^n \left[\text{length of } \frac{dc}{dt}(\xi_i) \right] \cdot (t_i - t_{i-1})$$

is the length of c , and the limit of such sums, for a general c , can be used as a definition of the length of c . The line integral

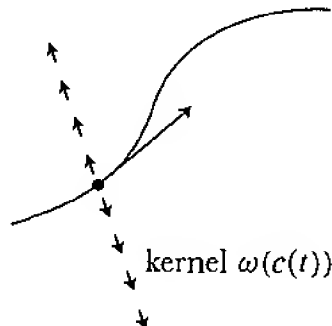
$$\int_c \omega = \text{limit of the sums } (*)$$

can be thought of as the “length” of c , when our ruler is changing continuously in a way specified by ω : Notice that the restriction of $\omega(c(t))$ to the 1-dimensional subspace of $\mathbb{R}^2_{c(t)}$ spanned by dc/dt is a constant times “signed length”. The natural way to specify a continuously changing length along c is to specify a length on its tangent vectors; this is the modern counterpart of the classical conception, whereby the curve c is divided into infinitely small parts, the infinitely small piece at $c(t)$, with components dx, dy , having length $f(c(t)) dx + g(c(t)) dy$.

Before pushing this geometrical interpretation too far, we should note that there is no 1-form ω on \mathbb{R}^2 such that

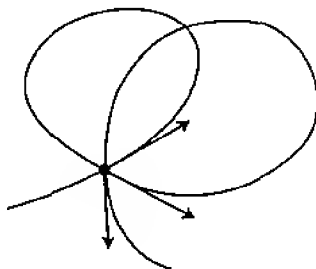
$$\int_c \omega = \text{length of } c \quad \text{for all curves } c.$$

It is true that for a given one-one curve c we can produce a form ω which works for c ; we choose $\omega(c(t)) \in \Omega^1(\mathbb{R}^2_{c(t)})$ so that

$$\omega(c(t)) \left(\frac{dc}{dt} \right) = 1,$$


(choosing the kernel of ω arbitrarily), and then extend ω to \mathbb{R}^2 . But if c is

not one-one this may be impossible; for example, in the situation shown below,



there is no element of $\Omega^1(\mathbb{R}^2_{c(t)})$ which has the value 1 on all three vectors. In general, given any ω on \mathbb{R}^2 which is everywhere non-zero, the subspaces $\Delta_p = \ker \omega(p)$ form a 1-dimensional distribution on \mathbb{R}^2 ; any curve contained in an integral submanifold of Δ will have “length” 0. Later we will see a way of circumventing this difficulty, if we are interested in obtaining the ordinary length of a curve. For the present, we note that the sums (*), used to define this generalized “length”, make sense even if c is a curve in a manifold M (where there is no notion of “length”), and ω is a 1-form on M , so we can define $\int_c \omega$ as the limit of these sums.

One property of line integrals should be mentioned now, because it is obvious with our original definition and merely true for our new definition. If $p: [0, 1] \rightarrow [0, 1]$ is a one-one increasing function from $[0, 1]$ onto $[0, 1]$, then the curve $c \circ p: [0, 1] \rightarrow M$ is called a **reparameterization** of c —it has exactly the same image as c , but transverses it at a different rate. Every sum $S(P, \xi)$ for c is clearly equal to a sum $S(P', \xi')$ for $c \circ p$, and conversely, so it is clear from our first definition that for a curve $c: [0, 1] \rightarrow \mathbb{R}^2$ we have

$$\int_c \omega = \int_{c \circ p} \omega$$

(“the integral of ω over c is independent of the parameterization”). This is no longer so clear when we consider the sums (*) for a curve $c: [0, 1] \rightarrow M$, nor is it clear even for a curve $c: [0, 1] \rightarrow \mathbb{R}^2$, but in this case we can proceed right to the integral these sums approach, namely

$$\int_0^1 [f(c(t))c^1'(t) + g(c(t))c^2'(t)] dt.$$

The result then follows from a calculation: the substitution $t = p(u)$ gives

$$\begin{aligned} & \int_0^1 [f(c(t))c'(t) + g(c(t))c^2'(t)] dt \\ &= \int_{p^{-1}(0)}^{p^{-1}(1)} [f(c(p(u)))c'(p(u)) + g(c(p(u)))c^2'(p(u))] p'(u) du \\ &= \int_0^1 [f(c \circ p(u))(c \circ p)'(u) + g(c \circ p(u))(c \circ p)^2'(u)] du. \end{aligned}$$

For a curve in \mathbb{R}^n , and a 1-form $\omega = \sum_{i=1}^n \omega_i dx^i$, there is a similar calculation; for a general manifold M , we can introduce a coordinate system for our calculations if $c([0, 1])$ lies in one coordinate system, or break c up into several pieces otherwise. We are being a bit sloppy about all this because we are about to introduce yet a third definition, which will eventually become our formal choice. Consider once again the case of a 1-form on \mathbb{R}^2 , where

$$\int_c \omega = \int_0^1 [f(c(t))c'(t) + g(c(t))c^2'(t)] dt.$$

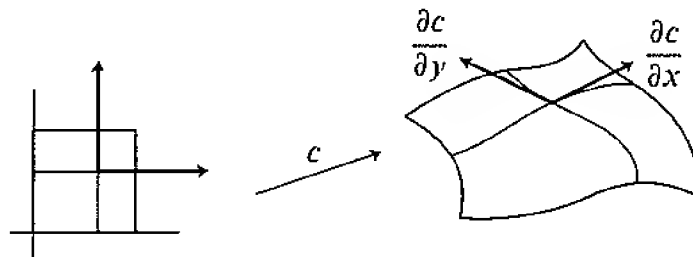
Notice that if t is the standard coordinate system on \mathbb{R} , then for the map $c: [0, 1] \rightarrow \mathbb{R}^2$ we have

$$\begin{aligned} c^*(f dx + g dy) &= (f \circ c)c^*(dx) + (g \circ c)c^*(dy) \\ &= (f \circ c)d(x \circ c) + (g \circ c)d(y \circ c) \\ &= (f \circ c)c' dt + (g \circ c)c^2' dt, \end{aligned}$$

so that formally we just integrate $c^*(f dx + g dy)$; to be precise, we write $c^*(f dx + g dy) = h dt$ (in the unique possible way), and take the integral of h on $[0, 1]$.

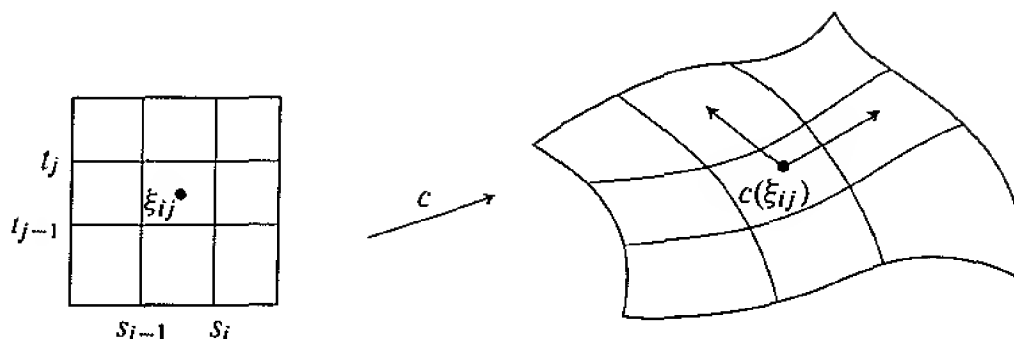
Everything we have said for curves $c: [0, 1] \rightarrow \mathbb{R}^n$ could be generalized to functions $c: [0, 1]^2 \rightarrow \mathbb{R}^n$. If x and y are the coordinate functions on \mathbb{R}^2 , let

$$\begin{aligned} \frac{\partial c}{\partial x} &= c_* \left(\frac{\partial}{\partial x} \right) \\ \frac{\partial c}{\partial y} &= c_* \left(\frac{\partial}{\partial y} \right). \end{aligned}$$



For a pair of partitions $s_0 < \dots < s_m$ and $t_0 < \dots < t_n$ of $[0, 1]$, if we choose

$\xi_{ij} \in [s_{i-1}, s_i] \times [t_{j-1}, t_j]$ and ω is a 2-form on \mathbb{R}^n , then



$$\omega(c(\xi_{ij})) \left(\frac{\partial c}{\partial x}(\xi_{ij}), \frac{\partial c}{\partial y}(\xi_{ij}) \right) (s_i - s_{i-1})(t_j - t_{j-1})$$

is a “generalized area” of the parallelogram spanned by

$$\frac{\partial c}{\partial x}(\xi_{ij}), \quad \frac{\partial c}{\partial y}(\xi_{ij}).$$

The limit of sums of these terms can be thought of as a “generalized area” of c . To make a long story short, we now proceed with the formal definitions.

A C^∞ function $c: [0, 1]^k \rightarrow M$ is called a **singular k -cube** in M (the word “singular” indicates that c is not necessarily one-one). We will let $[0, 1]^0 = \mathbb{R}^0 = 0 \in \mathbb{R}$, so that a singular 0-cube c is determined by the one point $c(0) \in M$. The inclusion map of $[0, 1]^k$ in \mathbb{R}^k will be denoted by $I^k: [0, 1]^k \rightarrow \mathbb{R}^k$; it is called the **standard k -cube**.

If ω is a k -form on $[0, 1]^k$, and x^1, \dots, x^k are the coordinate functions, then ω can be written uniquely as

$$\omega = f dx^1 \wedge \dots \wedge dx^k.$$

We define

$$\int_{[0,1]^k} \omega \quad \text{to be} \quad \int_{[0,1]^k} f \quad \left(\begin{array}{l} = \int_{[0,1]^k} f(x^1, \dots, x^k) dx^1 \dots dx^k \\ \text{in classical notation, which modern} \\ \text{notation attempts to mimic as far} \\ \text{as logic permits} \end{array} \right).$$

If ω is a k -form on M , and c is a singular k -cube in M , we define

$$\int_c \omega = \int_{[0,1]^k} c^* \omega,$$

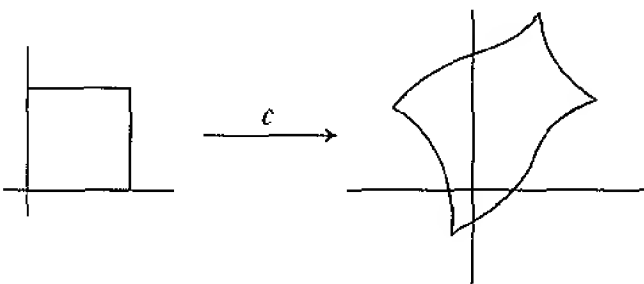
where the right hand side has just been defined. For $k = 0$, we have a special definition: a 0-form is a function f , and for a singular 0-cube c we define

$$\int_c f = f(c(0)).$$

1. PROPOSITION. Let $c: [0, 1]^n \rightarrow \mathbb{R}^n$ be a one-one singular n -cube with $\det c' \geq 0$ on $[0, 1]^n$. Let ω be the n -form

$$\omega = f dx^1 \wedge \cdots \wedge dx^n.$$

Then

$$\int_c \omega = \int_{c([0,1]^n)} f.$$


PROOF. By definition,

$$\begin{aligned} \int_c \omega &= \int_{[0,1]^n} c^*(\omega) \\ &= \int_{[0,1]^n} (f \circ c)(\det c') dx^1 \wedge \cdots \wedge dx^n \quad \text{by Theorem 7-7} \\ &= \int_{[0,1]^n} (f \circ c)|\det c'| dx^1 \wedge \cdots \wedge dx^n \quad \text{by assumption} \\ &= \int_{c([0,1]^n)} f \quad \text{by the change of variable formula. } \spadesuit \end{aligned}$$

2. COROLLARY. Let $p: [0, 1]^k \rightarrow [0, 1]^k$ be one-one onto with $\det p' \geq 0$, let c be a singular k -cube in M and let ω be a k -form on M . Then

$$\int_c \omega = \int_{c \circ p} \omega.$$

PROOF. We have

$$\begin{aligned} \int_{c \circ p} \omega &= \int_{[0,1]^k} (c \circ p)^* \omega = \int_{[0,1]^k} p^*(c^* \omega) \\ &= \int_{[0,1]^k} c^*(\omega) \quad \text{by the Proposition, since } p \text{ is onto} \\ &= \int_c \omega. \quad \spadesuit \end{aligned}$$

The map $c \circ p: [0, 1]^k \rightarrow M$ is called a **reparameterization** of c if $p: [0, 1]^k \rightarrow [0, 1]^k$ is a C^∞ one-one onto map with $\det p' \neq 0$ everywhere (so that p^{-1} is also C^∞); it is called **orientation preserving** or **orientation reversing** depending on whether $\det p' > 0$ or $\det p' < 0$ everywhere. The corollary thus shows independence of parameterization, provided it is orientation preserving; an orientation reversing reparameterization clearly changes the sign of the integral. Notice that there would be no such result if we tried to define the integral over c of a C^∞ function $f: M \rightarrow \mathbb{R}$ by the formula

$$\int_{[0,1]^k} f \circ c.$$

For example, if $c: [0, 1] \rightarrow M$ then

$$\int_0^1 f(c(t)) dt \quad \text{is generally} \quad \neq \int_0^1 f(c(p(t))) dt.$$

From a formal point of view, differential forms are the things we integrate because they transform correctly (i.e., in accordance with Theorem 7-7, so that the change of variable formula will pop up); functions on a manifold cannot be integrated (we can integrate a function f on the manifold \mathbb{R}^k only because it gives us a form $f dx^1 \wedge \cdots \wedge dx^k$).

Our definition of the integral of a k -form ω over a singular k -cube c can immediately be generalized. A k -chain is simply a formal (finite) sum of singular k -cubes multiplied by integers, e.g.,

$$1c_1 - 2c_2 + 3c_3.$$

The k -chain $1c_1 = 1 \cdot c_1$ will also be denoted simply by c_1 . We add k -chains, and multiply them by integers, purely formally, e.g.,

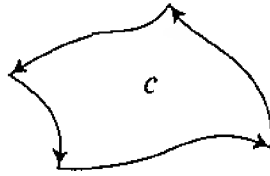
$$2(c_1 + 3c_4) + (-2)(c_1 + c_3 + c_2) = -2c_2 - 2c_3 + 6c_4.$$

Moreover, we define the integral of ω over a k -chain $c = \sum_i a_i c_i$ in the obvious way:

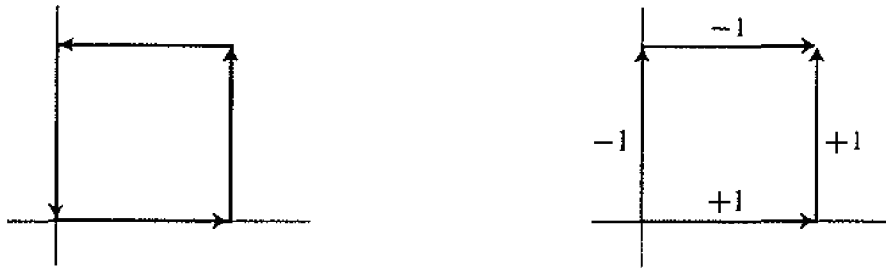
$$\int_{\sum_i a_i c_i} \omega = \sum_i a_i \int_{c_i} \omega.$$

The reason for introducing k -chains is that to every k -chain c (which may be just a singular k -cube) we wish to associate a $(k-1)$ -chain ∂c , which is called the **boundary** of c , and which is supposed to be the sum of the various singular

$(k - 1)$ -cubes around the boundary of each singular k -cube in c . In practice, it

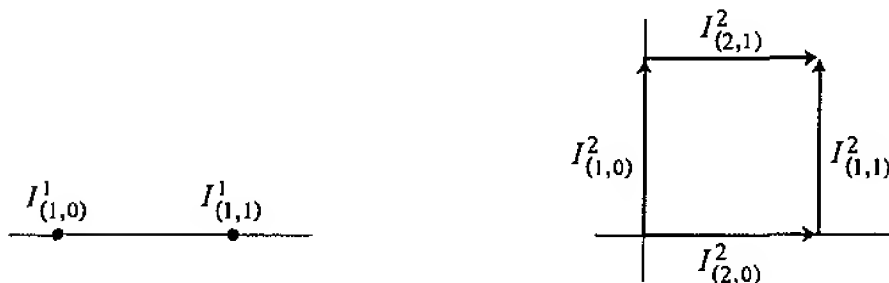


is convenient to modify this idea. The boundary of I^2 , for example, will not be the sum of the four singular 1-cubes indicated below on the left, but the sum,



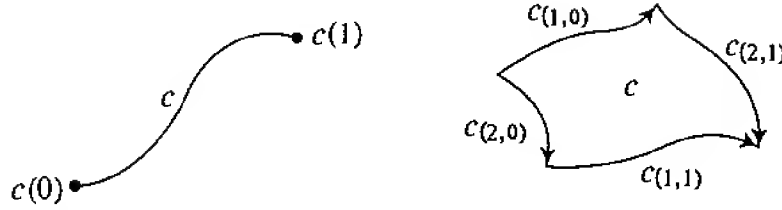
with the indicated coefficients, of the four singular 1-cubes shown on the right. (Notice that this will not change the integral of a 1-form over ∂I^2 .) For each i with $1 \leq i \leq n$ we first define two singular $(n - 1)$ -cubes $I_{(i,0)}^n$ and $I_{(i,1)}^n$ (the $(i, 0)$ -face and $(i, 1)$ -face of I^n) as follows: If $x \in [0, 1]^{n-1}$, then

$$\begin{aligned} I_{(i,0)}^n(x) &= I^n(x^1, \dots, x^{i-1}, 0, x^i, \dots, x^{n-1}) \\ &= (x^1, \dots, x^{i-1}, 0, x^i, \dots, x^{n-1}), \\ I_{(i,1)}^n(x) &= I^n(x^1, \dots, x^{i-1}, 1, x^i, \dots, x^{n-1}) \\ &= (x^1, \dots, x^{i-1}, 1, x^i, \dots, x^{n-1}). \end{aligned}$$



The (i, α) -face of a singular n -cube c is defined by

$$c_{(i,\alpha)} = c \circ (I^n_{(i,\alpha)}).$$



Now we define

$$\partial c = \sum_{i=1}^n \sum_{\alpha=0,1} (-1)^{i+\alpha} c_{(i,\alpha)}.$$

Finally, the boundary of an n -chain $\sum_i a_i c_i$ is defined by

$$\partial \left(\sum_i a_i c_i \right) = \sum_i a_i \partial(c_i).$$

These definitions all make sense only for $n \geq 1$. For the case of a 0-cube $c: [0, 1]^0 \rightarrow M$, which we will usually simply identify with the point $P = c(0)$, we define ∂c to be the number $1 \in \mathbb{R}$, and for a 0-chain $\sum_i a_i c_i$ we define

$$\partial \left(\sum_i a_i c_i \right) = \sum_i a_i \partial(c_i) = \sum_i a_i.$$

Notice that for a 1-cube $c: [0, 1] \rightarrow M$ we have

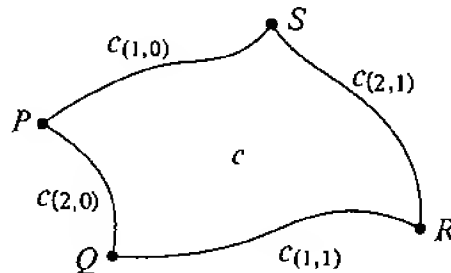
$$\partial c = c_{(1,1)} - c_{(1,0)},$$

so

$$\partial(\partial c) = 1 - 1 = 0.$$

We also have, for a singular 2-cube $c: [0, 1]^2 \rightarrow M$,

$$\begin{aligned} \partial c &= c_{(1,1)} - c_{(2,1)} - c_{(1,0)} + c_{(2,0)}, \\ \partial(\partial c) &= (R - Q) - (R - S) \\ &\quad - (S - P) + (Q - P) \\ &= 0. \end{aligned}$$



From a picture it can be checked that this also happens for a singular 3-cube, a good exercise because this involves figuring out just what the boundary of a 3-cube looks like. In general, we have:

3. PROPOSITION. If c is any n -chain in M , then $\partial(\partial c) = 0$. Briefly, $\partial^2 = 0$.

PROOF. Let $i \leq j \leq n-1$, and consider $(I_{(i,\alpha)}^n)_{(j,\beta)}$. For $x \in [0, 1]^{n-2}$, we have, from the definition

$$\begin{aligned} (I_{(i,\alpha)}^n)_{(j,\beta)}(x) &= I_{(i,\alpha)}^n(I_{(j,\beta)}^{n-1}(x)) \\ &= I_{(i,\alpha)}^n(x^1, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}) \\ &= I^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}). \end{aligned}$$

Similarly,

$$\begin{aligned} (I_{(j+1,\beta)}^n)_{(i,\alpha)} &= I_{(j+1,\beta)}^n(I_{(i,\alpha)}^{n-1}(x)) \\ &= I_{(j+1,\beta)}^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{n-2}) \\ &= I^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}). \end{aligned}$$

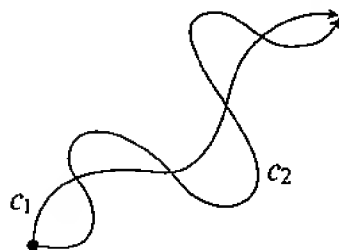
Thus $(I_{(i,\alpha)}^n)_{(j,\beta)} = (I_{(j+1,\beta)}^n)_{(i,\alpha)}$ for $i \leq j \leq n-1$. It follows easily for any singular n -cube c that $(c_{(i,\alpha)})_{(j,\beta)} = (c_{(j+1,\beta)})_{(i,\alpha)}$ for $i \leq j \leq n-1$. Now

$$\begin{aligned} \partial(\partial c) &= \partial \left(\sum_{i=1}^n \sum_{\alpha=0,1} (-1)^{i+\alpha} c_{(i,\alpha)} \right) \\ &= \sum_{i=1}^n \sum_{\alpha=0,1} \sum_{j=1}^{n-1} \sum_{\beta=0,1} (-1)^{i+\alpha+j+\beta} (c_{(i,\alpha)})_{(j,\beta)}. \end{aligned}$$

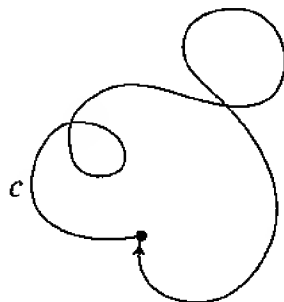
In this sum, $(c_{(i,\alpha)})_{(j,\beta)}$ and $(c_{(j+1,\beta)})_{(i,\alpha)}$ occur with opposite signs. Therefore all terms cancel in pairs, and $\partial(\partial c) = 0$. Since the theorem is true for singular n -cubes, it is clearly also true for singular n -chains. ♦

Notice that for some n -chains c we have not only $\partial(\partial c) = 0$, but even $\partial c = 0$. For example, this is the case if $c = c_1 - c_2$, where c_1 and c_2 are two 1-cubes

with $c_1(0) = c_2(0)$ and $c_1(1) = c_2(1)$. If c is just a singular 1-cube itself, then

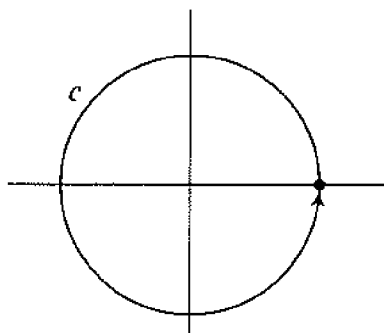


$\partial c = 0$ precisely when $c(0) = c(1)$, i.e., when c is a “closed” curve. In general,



any k -chain c is called closed if $\partial c = 0$.

Recall that a differential form ω with $d\omega = 0$ is also called “closed”; this terminology has been purposely chosen to parallel the terminology for chains (on the other hand, a chain of the form ∂c is not described, reciprocally, by the classical term of “exact”, but is simply called “a boundary”). This parallel terminology was not chosen merely because of the formal similarities between d and ∂ , expressed by the relations $d^2 = 0$ and $\partial^2 = 0$. The connection between forms and chains goes much deeper than that. For example, we have seen that on $\mathbb{R}^2 - \{0\}$ there is a 1-form “ $d\theta$ ” which is closed but not exact. There is also a 1-chain c which is closed but not a boundary, namely, a closed curve encircling



the point 0 once. Although it is intuitively clear that c is not the boundary of a 2-chain in $\mathbb{R}^2 - \{0\}$, the simplest proof uses the theorem which establishes the connection between forms, chains, d , and ∂ .

4. THEOREM (STOKES' THEOREM). If ω is a $(k-1)$ -form on M and c is a k -chain in M , then

$$\int_c d\omega = \int_{\partial c} \omega.$$

PROOF. Most of the proof involves the special case where ω is a $(k-1)$ -form on \mathbb{R}^k and $c = I^k$. In this case, ω is a sum of $(k-1)$ -forms of the type

$$f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k,$$

and it suffices to prove the theorem for each of these. We now compute. First, a little notation translation shows that

$$\begin{aligned} \int_{[0,1]^{k-1}} I_{(j,\alpha)}^k (f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\ = \begin{cases} 0 & \text{if } j \neq i \\ \int_{[0,1]^k} f(x^1, \dots, \alpha, \dots, x^k) dx^1 \dots dx^k & \text{if } j = i. \end{cases} \end{aligned}$$

Therefore

$$\begin{aligned} \int_{\partial I^k} f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k \\ = \sum_{j=1}^k \sum_{\alpha=0,1} (-1)^{j+\alpha} \int_{[0,1]^{k-1}} I_{(j,\alpha)}^k (f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\ = (-1)^{i+1} \int_{[0,1]^k} f(x^1, \dots, 1, \dots, x^k) dx^1 \dots dx^k \\ + (-1)^i \int_{[0,1]^k} f(x^1, \dots, 0, \dots, x^k) dx^1 \dots dx^k. \end{aligned}$$

On the other hand,

$$\begin{aligned}
 \int_{I^k} d(f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\
 &= \int_{[0,1]^k} D_i f dx^i \wedge dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k \\
 &= (-1)^{i-1} \int_{[0,1]^k} D_i f.
 \end{aligned}$$

By Fubini's theorem and the fundamental theorem of calculus we have

$$\begin{aligned}
 \int_{I^k} d(f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\
 &= (-1)^{i-1} \int_0^1 \cdots \left(\int_0^1 D_i f(x^1, \dots, x^k) dx^i \right) dx^1 \cdots \widehat{dx^i} \cdots dx^k \\
 &= (-1)^{i-1} \int_0^1 \cdots \int_0^1 \left[f(x^1, \dots, 1, \dots, x^k) \right. \\
 &\quad \left. - f(x^1, \dots, 0, \dots, x^k) \right] dx^1 \cdots \widehat{dx^i} \cdots dx^k \\
 &= (-1)^{i-1} \int_{[0,1]^k} f(x^1, \dots, 1, \dots, x^k) dx^1 \cdots dx^k \\
 &\quad + (-1)^i \int_{[0,1]^k} f(x^1, \dots, 0, \dots, x^k) dx^1 \cdots dx^k.
 \end{aligned}$$

Thus

$$\int_{I^k} d\omega = \int_{\partial I^k} \omega.$$

For an arbitrary singular k -cube, chasing through the definitions shows that

$$\int_{\partial c} \omega = \int_{\partial I^k} c^* \omega.$$

Therefore

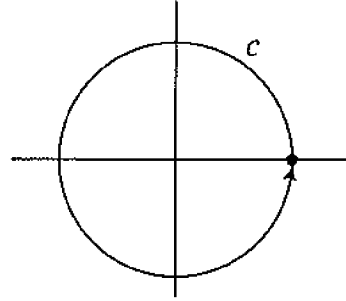
$$\int_c d\omega = \int_{I^k} c^*(d\omega) = \int_{I^k} d(c^*\omega) = \int_{\partial I^k} c^*\omega = \int_{\partial c} \omega.$$

The theorem clearly follows for k -chains also. ♦

Notice that Stokes' Theorem not only uses the fundamental theorem of calculus, but actually becomes that theorem when $c = I^1$ and $\omega = f$.

As an application of Stokes' Theorem, we show that the curve $c: [0, 1] \rightarrow \mathbb{R}^2 - \{0\}$ defined by

$$c(t) = (\cos 2\pi t, \sin 2\pi t),$$



although closed, is not ∂c^2 for any 2-chain c^2 . If we did have $c = \partial c^2$, then we would have

$$\int_c d\theta = \int_{\partial c^2} d\theta = \int_{c^2} d(d\theta) = \int_{c^2} 0 = 0.$$

But a straightforward computation (which will be good for the soul) shows that

$$\int_c d\theta = \int_c \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy = 2\pi.$$

[There is also a non-computational argument, using the fact that “ $d\theta$ ” really is $d\theta$ for $\theta: \mathbb{R}^2 - ([0, \infty) \times \{0\}) \rightarrow \mathbb{R}$: We have

$$\int_{c|[\varepsilon, 1-\varepsilon]} d\theta = \theta(1-\varepsilon) - \theta(\varepsilon),$$

and $\theta(1-\varepsilon) - \theta(\varepsilon) \rightarrow 2\pi$ as $\varepsilon \rightarrow 0$.]

Although we used this calculation to show that c is not a boundary, we could just as well have used it to show that $\omega = “d\theta”$ is not exact. For, if we had $\omega = df$ for some C^∞ function $f: \mathbb{R}^2 - \{0\} \rightarrow \mathbb{R}$, then we would have

$$2\pi = \int_c \omega = \int_c df = \int_{\partial c} f = \int_0 f = 0.$$

We were previously able to give a simpler argument to show that “ $d\theta$ ” is not exact, but Stokes' Theorem is the tool which will enable us to deal with forms on $\mathbb{R}^n - \{0\}$. For example, we will eventually obtain a 2-form ω on $\mathbb{R}^3 - \{0\}$,

$$\omega = \frac{x dy \wedge dz - y dx \wedge dz + z dx \wedge dy}{(x^2 + y^2 + z^2)^{3/2}}$$

which is closed but not exact. For the moment we are keeping the origin of ω a secret, but a straightforward calculation shows that $d\omega = 0$. To prove that ω is not exact we will want to integrate it over a 2-chain which “fills up” the 2-sphere $S^2 \subset \mathbb{R}^3 - \{0\}$. There are lots of ways of doing this, but they all turn out to give the same result. In fact, we first want to describe a way of integrating n -forms over n -manifolds. This is possible only when M is orientable; the reason will be clear from the next result, which is basic for our definition.

5. THEOREM. Let M be an n -manifold with an orientation μ , and let $c_1, c_2 : [0, 1]^n \rightarrow M$ be two singular n -cubes which can be extended to be diffeomorphisms in a neighborhood of $[0, 1]^n$. Assume that c_1 and c_2 are both *orientation preserving* (with respect to the orientation μ on M , and the usual orientation on \mathbb{R}^n). If ω is an n -form on M such that

$$\text{support } \omega \subset c_1([0, 1]^n) \cap c_2([0, 1]^n),$$

then

$$\int_{c_1} \omega = \int_{c_2} \omega.$$

PROOF. We want to use Corollary 2, and write

$$\int_{c_2} \omega = \int_{c_2 \circ (c_2^{-1} \circ c_1)} \omega = \int_{c_1} \omega.$$

The only problem is that $c_2^{-1} \circ c_1$ is not defined on all of $[0, 1]^n$ (it does satisfy $\det(c_2^{-1} \circ c_1)' \geq 0$, since c_1 and c_2 are both orientation preserving). However, a glance at the proof of Corollary 2 will show that the result still follows, because of the fact that $\text{support } \omega \subset c_1([0, 1]^n) \cap c_2([0, 1]^n)$. ♦

The common number $\int_c \omega$, for singular n -cubes $c : [0, 1]^n \rightarrow M$ with $\text{support } \omega \subset c([0, 1]^n)$ and c orientation preserving, will be denoted by

$$\int_M \omega.$$

If ω is an arbitrary n -form on M , then there is a cover \mathcal{O} of M by open sets U , each contained in some $c([0, 1]^n)$, where c is a singular n -cube of this sort; if Φ is a partition of unity subordinate to this cover, then

$$\int_M \phi \cdot \omega$$

is defined for each $\phi \in \Phi$. We wish to define

$$\int_M \omega = \sum_{\phi \in \Phi} \int_M \phi \cdot \omega.$$

We will adopt this definition only when ω has compact support, in which case the sum is actually finite, since support ω can intersect only finitely many of the sets $\{p : \phi(p) \neq 0\}$, which form a locally finite collection. If we have another partition of unity Ψ (subordinate to a cover \mathcal{O}'), then

$$\sum_{\phi \in \Phi} \int_M \phi \cdot \omega = \sum_{\phi \in \Phi} \int_M \sum_{\psi \in \Psi} \psi \cdot \phi \cdot \omega = \sum_{\phi \in \Phi} \sum_{\psi \in \Psi} \int_M \psi \cdot \phi \cdot \omega;$$

these sums are all finite, and the last sum can clearly also be written as

$$\sum_{\psi \in \Psi} \sum_{\phi \in \Phi} \int_M \phi \cdot \psi \cdot \omega = \sum_{\psi \in \Psi} \int_M \psi \cdot \omega,$$

so that our definition does not depend on the partition. (We really should denote this sum by

$$\int_{(M, \mu)} \omega;$$

for the orientation $-\mu$ of M we clearly have

$$\int_{(M, -\mu)} \omega = - \int_{(M, \mu)} \omega.$$

However, we usually omit explicit mention of μ .)

With minor modifications we can define $\int_M \omega$ even if M is an n -manifold-with-boundary. If $M \subset \mathbb{R}^n$ is an n -dimensional manifold-with-boundary and $f: M \rightarrow \mathbb{R}$ has compact support, then

$$\int_M f dx^1 \wedge \cdots \wedge dx^n = \int_M f.$$

where the right hand side denotes the ordinary integral. This is a simple consequence of Proposition 1. Likewise, if $f: M^n \rightarrow N^n$ is a diffeomorphism onto, and ω is an n -form with compact support on N , then

$$\int_M f^* \omega = \begin{cases} \int_N \omega & \text{if } f \text{ is orientation preserving} \\ - \int_N \omega & \text{if } f \text{ is orientation reversing.} \end{cases}$$

Although n -forms can be integrated only over orientable manifolds, there is a way of discussing integration on non-orientable manifolds. Suppose that ω is a function on M such that for each $p \in M$ we have

$$\omega(p) = |\eta_p| \quad \text{for some } \eta_p \in \Omega^n(M_p),$$

i.e., for any n vectors $v_1, \dots, v_n \in M_p$ we have

$$\omega(p)(v_1, \dots, v_n) = |\eta_p(v_1, \dots, v_n)| \geq 0.$$

Such a function ω is called a **volume element**—on each vector space it determines a way of measuring n -dimensional volume (not signed volume). If (x, U) is a coordinate system, then on U we can write

$$\omega = f |dx^1 \wedge \dots \wedge dx^n| \quad \text{for } f \geq 0;$$

we call ω a C^∞ volume element if f is C^∞ . One way of obtaining a volume element is to begin with an n -form η and then define $\omega(p) = |\eta(p)|$. However, not every volume element arises in this way—the form η_p may not vary continuously with p . For example, consider the Möbius strip M , imbedded in \mathbb{R}^3 . Since M_p can be considered as a subspace of \mathbb{R}^3_p , we can define

$$\omega(p)(v_p, w_p) = \text{area of parallelogram spanned by } v \text{ and } w.$$

It is not hard to see that ω is a volume element; locally, ω is of the form $\omega = |\eta|$ for an n -form η . But this cannot be true on all of M , since there is no n -form η on M which is everywhere non-zero.

Theorem 7-7 has an obvious modification for volume elements:

7-7'. THEOREM. If $f: M \rightarrow N$ is a C^∞ function between n -manifolds, (x, U) is a coordinate system around $p \in M$, and (y, V) a coordinate system around $q = f(p) \in N$, then for non-negative $g: V \rightarrow \mathbb{R}$ we have

$$f^*(g |dy^1 \wedge \dots \wedge dy^n|) = (g \circ f) \cdot \left| \det \left(\frac{\partial(y^i \circ f)}{\partial x^j} \right) \right| \cdot |dx^1 \wedge \dots \wedge dx^n|.$$

PROOF. Go through the proof of Theorem 7-7, putting in absolute value signs in the right place. ♦

7-8'. COROLLARY. If (x, U) and (y, V) are two coordinate systems on M and

$$g |dy^1 \wedge \cdots \wedge dy^n| = h |dx^1 \wedge \cdots \wedge dx^n| \quad g, h \geq 0$$

then

$$h = g \cdot \left| \det \left(\frac{\partial y^i}{\partial x^j} \right) \right|.$$

[This corollary shows that volume elements are the geometric objects corresponding to the "odd scalar densities" defined in Problem 4-10.]

It is now an easy matter to integrate a volume element ω over any manifold. First we define

$$\int_{[0,1]^n} \omega = \int_{[0,1]^n} f \quad \text{for } \omega = f |dx^1 \wedge \cdots \wedge dx^n|, \quad f \geq 0.$$

Then for an n -chain $c: [0, 1]^n \rightarrow M$ we define

$$\int_c \omega = \int_{[0,1]^n} c^* \omega.$$

Theorem 7-7' shows that Proposition 1 holds for a volume element $\omega = f |dx^1 \wedge \cdots \wedge dx^n|$ even if $\det c'$ is not ≥ 0 . Thus Corollary 2 holds for volume elements even if $\det p'$ is not ≥ 0 . From this we conclude that Theorem 5 holds for volume elements ω on any manifold M , without assuming c_1, c_2 orientation preserving (or even that M is orientable). Consequently we can define $\int_M \omega$ for any volume element ω with compact support.

Of course, when M is orientable these considerations are unnecessary. For, there is a nowhere zero n -form η on M , and consequently any volume element ω can be written

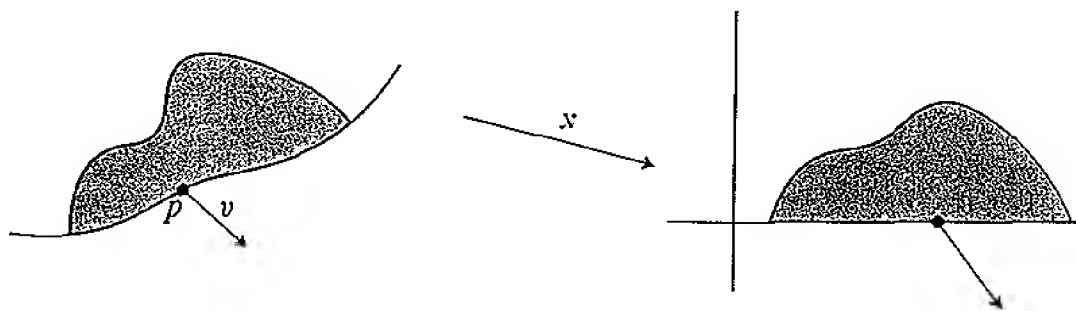
$$\omega = f |\eta|, \quad f \geq 0.$$

If we choose an orientation μ for M such that $\omega(v_1, \dots, v_n) > 0$ for v_1, \dots, v_n positively oriented, then we can define

$$\int_M \omega = \int_{(M, \mu)} f \eta.$$

Volume elements will be important later, but for the remainder of this chapter we are concerned only with integrating forms over oriented manifolds. In fact, our main result about integrals of forms over manifolds, an analogue of Stokes' Theorem about the integral of forms over chains, does not work for volume elements.

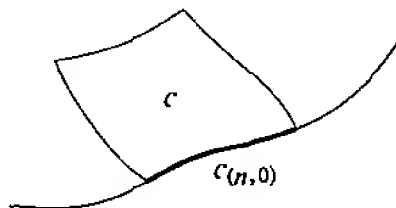
Recall from Problem 3-16 that if M is a manifold-with-boundary, and $p \in \partial M$, then certain vectors $v \in M_p$ can be distinguished by the fact that for any coordinate system $x: U \rightarrow \mathbb{H}^n$ around p , the vector $x_*(v) \in \mathbb{H}^n_{f(p)}$ points “outwards”. We call such vectors $v \in M_p$ “outward pointing”. If M has an



orientation μ , we define the induced orientation $\partial\mu$ for ∂M by the condition that $[v_1, \dots, v_{n-1}] \in (\partial\mu)_p$ if and only if $[w, v_1, \dots, v_{n-1}] \in \mu_p$ for every outward pointing $w \in M_p$. If μ is the usual orientation of \mathbb{H}^n , then for $p = (a, 0) \in \mathbb{H}^n$ we have

$$\begin{aligned}\mu_p &= [(e_1)_p, \dots, (e_n)_p] = (-1)^{n-1}[(e_n)_p, (e_1)_p, \dots, (e_{n-1})_p] \\ &= (-1)^n[(-e_n)_p, (e_1)_p, \dots, (e_{n-1})_p].\end{aligned}$$

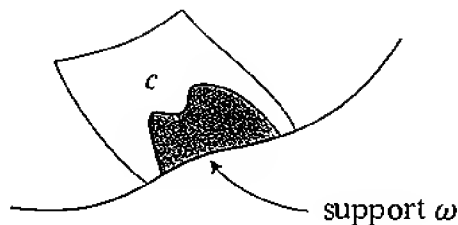
Since $(-e_n)_p$ is an outward pointing vector, this shows that the induced orientation on $\mathbb{R}^{n-1} \times \{0\} = \partial\mathbb{H}^n$ is $(-1)^n$ times the usual one. The reason for this choice is the following. Let c be an orientation preserving singular n -cube in (M, μ) such that $\partial M \cap c([0, 1]^n) = c_{(n,0)}([0, 1]^{n-1})$. Then $c_{(n,0)}: [0, 1]^{n-1} \rightarrow$



$(\partial M, \partial\mu)$ is orientation preserving for even n , and orientation reversing for odd n . If ω is an $(n-1)$ -form on M whose support is contained in the interior

of the image of c (this interior contains points in the image of $c_{(n,0)}$), it follows that

$$\int_{c_{(n,0)}} \omega = (-1)^n \int_{\partial M} \omega.$$



But $c_{(n,0)}$ appears with coefficient $(-1)^n$ in ∂c . So

$$(*) \quad \int_{\partial c} \omega = \int_{(-1)^n c_{(n,0)}} \omega = (-1)^n \int_{c_{(n,0)}} \omega = \int_{\partial M} \omega.$$

If it were not for this choice of $\partial\mu$ we would have some unpleasant minus signs in the following theorem.

6. THEOREM (STOKES' THEOREM). If M is an oriented n -dimensional manifold-with-boundary, and ∂M is given the induced orientation, and ω is an $(n-1)$ -form on M with compact support, then

$$\int_M d\omega = \int_{\partial M} \omega.$$

PROOF. Suppose first that there is an orientation preserving singular n -cube c in $M - \partial M$ such that $\text{support } \omega \subset \text{interior of image } c$. Then

$$\begin{aligned} \int_M d\omega &= \int_c d\omega = \int_{\partial c} \omega && \text{by Theorem 4} \\ &= 0 && \text{since support } \omega \subset \text{interior of image } c, \end{aligned}$$

while we clearly have

$$\int_{\partial M} \omega = 0.$$

Suppose next that there is an orientation preserving singular n -cube c in M such that $\partial M \cap c([0, 1]^n) = c_{(n,0)}([0, 1]^{n-1})$, and $\text{support } \omega \subset \text{interior of image } c$. Then once again

$$\int_M d\omega = \int_c d\omega = \int_{\partial c} \omega = \int_{\partial M} \omega \quad \text{by } (*).$$

In general, there is an open cover \mathcal{O} of M and a partition of unity Φ subordinate to \mathcal{O} such that for each $\phi \in \Phi$ the form $\phi \cdot \omega$ is one of the two sorts already considered. We have

$$0 = d(1) = d\left(\sum_{\phi \in \Phi} \phi\right) = \sum_{\phi \in \Phi} d\phi,$$

so

$$\sum_{\phi \in \Phi} d\phi \wedge \omega = 0.$$

Since ω has compact support, this is really a finite sum, and we conclude that

$$\sum_{\phi \in \Phi} \int_M d\phi \wedge \omega = 0.$$

Therefore

$$\begin{aligned} \int_M d\omega &= \sum_{\phi \in \Phi} \int_M \phi \cdot d\omega = \sum_{\phi \in \Phi} \int_M d\phi \wedge \omega + \phi \cdot d\omega \\ &= \sum_{\phi \in \Phi} \int_M d(\phi \cdot \omega) = \sum_{\phi \in \Phi} \int_{\partial M} \phi \cdot \omega = \int_{\partial M} \omega. \quad \spadesuit \end{aligned}$$

One of the simplest applications of Stokes' Theorem occurs when the oriented n -manifold (M, μ) is compact (so that every form has compact support) and $\partial M = \emptyset$. In this case, if η is any $(n-1)$ -form, then

$$\int_M d\eta = \int_{\partial M} \eta = 0.$$

Therefore we can find an n -form ω on M which is *not* exact (even though it must be closed, because all $(n+1)$ -forms on M are 0), simply by finding an ω with

$$\int_M \omega \neq 0.$$

Such a form ω always exists. Indeed we have seen that there is a form ω such that for $v_1, \dots, v_n \in M_p$ we have

$$(*) \quad \omega(v_1, \dots, v_n) > 0 \quad \text{if } [v_1, \dots, v_n] = \mu_p.$$

If $c: [0, 1]^n \rightarrow (M, \mu)$ is orientation preserving, then the form $c^*\omega$ on $[0, 1]^n$ is clearly

$$g \, dx^1 \wedge \dots \wedge dx^n \quad \text{for some } g > 0 \text{ on } [0, 1]^n,$$

so $\int_c \omega > 0$. It follows that $\int_M \omega > 0$. There is, moreover, no need to choose a form ω with (*) holding everywhere—we can allow the $>$ sign to be replaced by \geq . Thus we can even obtain a non-exact n -form on M which has support contained in a coordinate neighborhood.

This seemingly minor result already proves a theorem: a compact oriented manifold is not smoothly contractible to a point. As we have already emphasized, it is the “shape” of M , rather than its “size”, which determines whether or not every closed form on M is exact. Roughly speaking, we can obtain more information about the shape of M by analyzing more closely the extent to which closed forms are not necessarily exact. In particular, we would now like to ask just how many non-exact n -forms there are on a compact oriented n -manifold M . Naturally, if ω is not exact, then the same is true for $\omega + d\eta$ for any $(n-1)$ -form η , so we really want to consider ω and $\omega + d\eta$ as equivalent. There is, of course, a standard way of doing this, by considering quotient spaces. We will apply this construction not only to n -forms, but to forms of any degree.

For each k , the collection $Z^k(M)$ of all closed k -forms on M is a vector space. The space $B^k(M)$ of all exact k -forms is a subspace (since $d^2 = 0$), so we can form the quotient vector space

$$H^k(M) = Z^k(M)/B^k(M);$$

this vector space $H^k(M)$ is called the k -dimensional de Rham cohomology vector space of M . [*de Rham's Theorem* states that this vector space is isomorphic to a certain vector space defined purely in terms of the topology of M (for any space M), called the “ k -dimensional cohomology group of M with real coefficients”; the notation Z^k , B^k is chosen to correspond to the notation used in algebraic topology, where these groups are defined.]

An element of $H^k(M)$ is an equivalence class $[\omega]$ of a closed k -form ω , two closed k -forms ω_1 and ω_2 being equivalent if and only if their difference is exact. In terms of these vector spaces, the Poincaré Lemma says that $H^k(\mathbb{R}^n) = 0$ (the vector space containing only 0) if $k > 0$, or more generally, $H^k(M) = 0$ if M is contractible and $k > 0$.

To compute $H^0(M)$ we note first that $B^0(M) = 0$ (there are no non-zero exact 0-forms, since there are no non-zero (-1) -forms for them to be the differential of). So $H^0(M)$ is the same as the vector space of all C^∞ functions $f: M \rightarrow \mathbb{R}$ with $df = 0$. If M is connected, the condition $df = 0$ implies that f is constant, so $H^0(M) \approx \mathbb{R}$. (In general, the dimension of $H^0(M)$ is the number of components of M .)

Aside from these trivial remarks, we presently know only one other fact about $H^k(M)$ —if M is compact and oriented, then $H^n(M)$ has dimension ≥ 1 . The further study of $H^k(M)$ requires a careful look at spheres and Euclidean space.

On $S^{n-1} \subset \mathbb{R}^n - \{0\}$ there is a natural choice of an $(n-1)$ -form σ' with $\int_{S^{n-1}} \sigma' > 0$: for $(v_1)_p, \dots, (v_{n-1})_p \in S^{n-1}_p$, we define

$$\sigma'(p)((v_1)_p, \dots, (v_{n-1})_p) = \det \begin{pmatrix} p \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix}.$$

Clearly this is > 0 if $(v_1)_p, \dots, (v_{n-1})_p$ is a positively oriented basis. In fact, we defined the orientation of S^{n-1} in precisely this way—this orientation is just the induced orientation when S^{n-1} is considered as the boundary of the unit ball $\{p \in \mathbb{R}^n : |p| \leq 1\}$ with the usual orientation. Using the expansion of a determinant by minors along the top row we see that σ' is the restriction to S^{n-1} of the form σ on \mathbb{R}^n defined by

$$\sigma = \sum_{i=1}^n (-1)^{i-1} x^i dx^1 \wedge \dots \wedge \widehat{dx^i} \wedge \dots \wedge dx^n.$$

The form σ' on S^{n-1} will now be used to find an $(n-1)$ -form on $\mathbb{R}^n - \{0\}$ which is closed but not exact (thus showing that $H^{n-1}(\mathbb{R}^n - \{0\}) \neq 0$). Consider the map $r: \mathbb{R}^n - \{0\} \rightarrow S^{n-1}$ defined by

$$r(p) = \frac{p}{|p|} = \frac{p}{v(p)}.$$

Clearly $r(p) = p$ if $p \in S^{n-1}$; otherwise said, if $i: S^{n-1} \rightarrow \mathbb{R}^n - \{0\}$ is the inclusion, then

$$r \circ i = \text{identity of } S^{n-1}.$$

(In general, if $A \subset X$ and $r: X \rightarrow A$ satisfies $r(a) = a$ for $a \in A$, then r is called a **retraction** of X onto A .)

Clearly, $r^*\sigma'$ is closed:

$$d(r^*\sigma') = r^*d\sigma' = 0.$$

However, it is not exact, for if $r^*\sigma' = d\eta$, then

$$\sigma' = i^*r^*\sigma' = di^*\eta;$$

but we know that σ' is not exact.

It is a worthwhile exercise to compute by brute force that

$$\text{for } n = 2, \quad r^*\sigma' = \frac{x dy - y dx}{x^2 + y^2} = \frac{x dy - y dx}{v^2} = d\theta$$

$$\begin{aligned} \text{for } n = 3, \quad r^*\sigma' &= \frac{x dy \wedge dz - y dx \wedge dz + z dx \wedge dy}{(x^2 + y^2 + z^2)^{3/2}} \\ &= \frac{1}{v^3} [x dy \wedge dz - y dx \wedge dz + z dx \wedge dy]. \end{aligned}$$

Since we will actually need to know $r^*\sigma'$ in general, we evaluate it in another way:

7. LEMMA. If σ is the form on \mathbb{R}^n defined by

$$\sigma = \sum_{i=1}^n (-1)^{i-1} x^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n,$$

and σ' is the restriction $i^*\sigma$ of σ to S^{n-1} , then

$$(*) \quad r^*\sigma'(p) = \frac{\sigma(p)}{|p|^n}.$$

So

$$r^*\sigma' = \frac{1}{v^n} \sum_{i=1}^n (-1)^{i-1} x^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n.$$

PROOF. At any point $p \in \mathbb{R}^n - \{0\}$, the tangent space \mathbb{R}_p^n is spanned by p_p and the vectors v_p in the tangent space of the sphere $S^{n-1}(|p|)$ of radius $|p|$. So it suffices to check that both sides of $(*)$ give the same result when applied to $n-1$ vectors each of which is one of these two sorts. Now p_p is the tangent vector of a curve γ lying along the straight line through 0 and p ; this curve is taken to the single point $r(p)$ by r , so $r_*(p_p) = 0$. On the other hand,

$$\sigma(p)(p_p, (v_1)_p, \dots, (v_{n-2})_p) = \det \begin{pmatrix} p \\ p \\ v_1 \\ \vdots \\ v_{n-2} \end{pmatrix} = 0.$$

So it suffices to apply both sides of $(*)$ to vectors in the tangent space of $S^{n-1}(|p|)$. Thus (Problem 15), it suffices to show that for such vectors v_p we have

$$r_*(v_p) = \frac{1}{|p|} v_{r(p)}.$$

But this is almost obvious, since the vector v_p is the tangent vector of a circle γ lying in $S^{n-1}(|p|)$, and the curve $r \circ \gamma$ lies in S^{n-1} and goes $1/|p|$ as far in the same time. ♦

8. COROLLARY (INTEGRATION IN "POLAR COORDINATES"). Let $f: B \rightarrow \mathbb{R}$, where

$$B = \{p \in \mathbb{R}^n : |p| \leq 1\},$$

and define $g: S^{n-1} \rightarrow \mathbb{R}$ by

$$g(p) = \int_0^1 u^{n-1} f(u \cdot p) du.$$

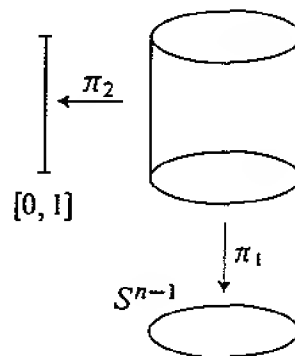
Then

$$\int_B f = \int_B f dx^1 \wedge \cdots \wedge dx^n = \int_{S^{n-1}} g \sigma'.$$

PROOF. Consider $S^{n-1} \times [0, 1]$ and the two projections

$$\pi_1: S^{n-1} \times [0, 1] \rightarrow S^{n-1}$$

$$\pi_2: S^{n-1} \times [0, 1] \rightarrow [0, 1].$$



Let us use the abbreviation

$$\sigma' \wedge dt = \pi_1^* \sigma' \wedge \pi_2^* dt.$$

If (y, U) is a coordinate system on S^{n-1} , with a corresponding coordinate system $(\bar{y}, t) = (y \circ \pi_1, \pi_2)$ on $S^{n-1} \times [0, 1]$, and $\sigma' = \alpha dy^1 \wedge \cdots \wedge dy^{n-1}$, then clearly

$$\sigma' \wedge dt = \bar{\alpha} \circ \pi_1 d\bar{y}^1 \wedge \cdots \wedge d\bar{y}^{n-1} \wedge dt.$$

From this it is easy to see that if we define $h: S^{n-1} \times [0, 1] \rightarrow \mathbb{R}$ by

$$h(p, u) = u^{n-1} f(u \cdot p),$$

then

$$\int_{S^{n-1}} g \sigma' = (-1)^{n-1} \int_{S^{n-1} \times [0, 1]} h \sigma' \wedge dt.$$

Now we can define a diffeomorphism $\phi: B - \{0\} \rightarrow S^{n-1} \times (0, 1]$ by

$$\phi(p) = (r(p), v(p)) = (p/|p|, |p|).$$

Then

$$\begin{aligned}
 \phi^*(\sigma' \wedge dt) &= \phi^*(\pi_1^*\sigma' \wedge \pi_2^*dt) \\
 &= \phi^*\pi_1^*\sigma' \wedge \phi^*\pi_2^*dt \\
 &= (\pi_1 \circ \phi)^*\sigma' \wedge (\pi_2 \circ \phi)^*dt \\
 &= r^*\sigma' \wedge v^*dt \\
 &= \frac{1}{v^n} \left(\sum_{i=1}^n (-1)^{i-1} x^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n \right) \wedge \sum_{i=1}^n \frac{x^i}{v} dx^i \\
 &= \frac{(-1)^{n-1}}{v^{n+1}} \sum_{i=1}^n (x^i)^2 dx^1 \wedge \cdots \wedge dx^n \\
 &= \frac{(-1)^{n-1}}{v^{n-1}} dx^1 \wedge \cdots \wedge dx^n.
 \end{aligned}$$

Hence

$$\begin{aligned}
 \phi^*(h\sigma' \wedge dt) &= (h \circ \phi)\phi^*(\sigma' \wedge dt) \\
 &= v^{n-1} f \cdot \frac{(-1)^{n-1}}{v^{n-1}} dx^1 \wedge \cdots \wedge dx^n \\
 &= (-1)^{n-1} f dx^1 \wedge \cdots \wedge dx^n.
 \end{aligned}$$

So,

$$\begin{aligned}
 \int_B f dx^1 \wedge \cdots \wedge dx^n &= (-1)^{n-1} \int_{B-\{0\}} \phi^*(h\sigma' \wedge dt) \\
 &= (-1)^{n-1} \int_{S^{n-1} \times (0,1]} h\sigma' \wedge dt \\
 &= \int_{S^{n-1}} g\sigma'.
 \end{aligned}$$

(This last step requires some justification, which should be supplied by the reader, since the forms involved do not have compact support on the manifolds $B - \{0\}$ and $S^{n-1} \times (0, 1]$ where they are defined.) ♦

We are about ready to compute $H^k(M)$ in a few more cases. We are going to reduce our calculations to calculations within coordinate neighborhoods, which are submanifolds of M , but not compact. It is therefore necessary to introduce another collection of vector spaces, which are interesting in their own right.

The de Rham cohomology vector spaces with compact supports $H_c^k(M)$ are defined as

$$H_c^k(M) = Z_c^k(M) / B_c^k(M),$$

where $Z_c^k(M)$ is the vector space of closed k -forms with compact support, and $B_c^k(M)$ is the vector space of all k -forms $d\eta$ where η is a $(k-1)$ -form with compact support. Of course, if M is compact, then $H_c^k(M) = H^k(M)$. Notice that $B_c^k(M)$ is *not* the same as the set of all exact k -forms with compact support. For example, on \mathbb{R}^n , if $f \geq 0$ is a function with compact support, and $f > 0$ at some point, then

$$\omega = f dx^1 \wedge \cdots \wedge dx^n$$

is exact (every closed form on \mathbb{R}^n is) and has compact support, but ω is not $d\eta$ for any form η with compact support. Indeed, if $\omega = d\eta$ where η has compact support, then by Stokes' Theorem

$$\int_{\mathbb{R}^n} \omega = \int_{\mathbb{R}^n} d\eta = \int_{\partial\mathbb{R}^n} \eta = 0.$$

This example shows that $H_c^n(\mathbb{R}^n) \neq 0$, and a similar argument shows that if M is any orientable manifold, then $H_c^n(M) \neq 0$. We are now going to show that for any connected orientable manifold M we actually have

$$H_c^n(M) \approx \mathbb{R}.$$

This means that if we choose a fixed ω with $\int_M \omega \neq 0$, then for any n -form ω' with compact support there is a real number a such that $\omega' - a\omega$ is exact. The number a can be described easily: if

$$\omega' - a\omega = d\eta,$$

then

$$\int_M \omega' - \int_M a\omega = \int_M d\eta = 0,$$

so

$$a = \int_M \omega' / \int_M \omega;$$

the problem, of course, is showing that η exists. Notice that the assertion that $H_c^n(M) \approx \mathbb{R}$ is equivalent to the assertion that

$$[\omega] \mapsto \int_M \omega$$

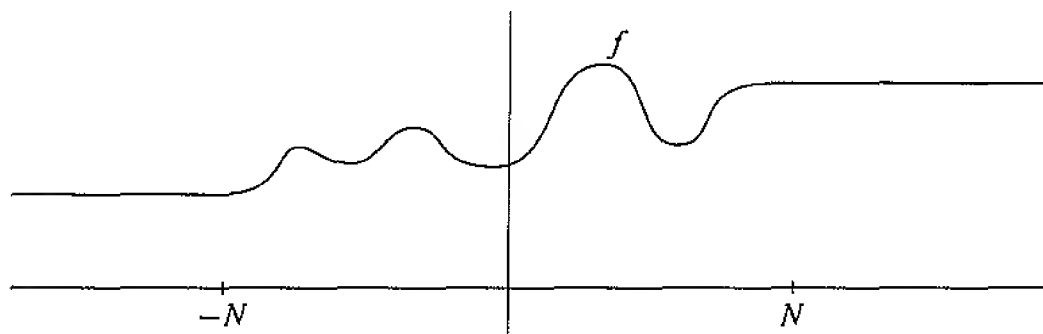
is an isomorphism of $H_c^n(M)$ with \mathbb{R} , i.e., to the assertion that a closed form ω with compact support is the differential of another form with compact support if $\int_M \omega = 0$.

9. THEOREM. If M is a connected orientable n -manifold, then $H_c^n(M) \approx \mathbb{R}$.

PROOF. We will establish the theorem in three steps:

- (1) The theorem is true for $M = \mathbb{R}$.
- (2) If the theorem is true for $(n-1)$ -manifolds, in particular for S^{n-1} , then it is true for \mathbb{R}^n .
- (3) If the theorem is true for \mathbb{R}^n , then it is true for any connected oriented n -manifold.

Step 1. Let ω be a 1-form on \mathbb{R} with compact support such that $\int_{\mathbb{R}} \omega = 0$. There is some function f (not necessarily with compact support) such that $\omega = df$. Since support ω is compact, $df = 0$ outside some interval $[-N, N]$, so f is a



constant c_1 on $(-\infty, -N)$ and a constant c_2 on (N, ∞) . Moreover,

$$0 = \int_{\mathbb{R}} \omega = \int_{\mathbb{R}} df = \int_{\mathbb{R}} f'(t) dt = c_2 - c_1.$$

Therefore $c_1 = c_2 = c$ and we have

$$\omega = d(f - c)$$

where $f - c$ has compact support.

Step 2. Let $\omega = f dx^1 \wedge \cdots \wedge dx^n$ be an n -form with compact support on \mathbb{R}^n such that $\int_{\mathbb{R}^n} \omega = 0$. For simplicity assume that support $\omega \subset \{p \in \mathbb{R}^n : |p| < 1\}$. We know that there is an $(n-1)$ -form η on \mathbb{R}^n such that $\omega = d\eta$. In fact, from Problem 7-23, we have an explicit formula for η ,

$$\eta(p) = \sum_{i=1}^n (-1)^{i-1} \left(\int_0^1 t^{n-1} f(t \cdot p) dt \right) x^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n.$$

Using the substitution $u = |p|t$ this becomes

$$\begin{aligned}\eta(p) &= \left(\int_0^{|p|} u^{n-1} f\left(u \cdot \frac{p}{|p|}\right) du \right) \frac{1}{|p|^n} \\ &\quad \times \sum_{i=1}^n (-1)^{i-1} x^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n \\ &= \left(\int_0^{|p|} u^{n-1} f\left(u \cdot \frac{p}{|p|}\right) du \right) \cdot r^* \sigma'(p) \quad \text{by Lemma 7.}\end{aligned}$$

Define $g: S^{n-1} \rightarrow \mathbb{R}$ by

$$g(p) = \int_0^1 u^{n-1} f(u \cdot p) du.$$

On the set $A = \{p \in \mathbb{R}^n : |p| > 1\}$ we have $f = 0$, so on A we have

$$\eta(p) = \left(\int_0^1 u^{n-1} f\left(u \cdot \frac{p}{|p|}\right) du \right) \cdot r^* \sigma'(p),$$

or

$$\eta = (g \circ r) \cdot r^* \sigma' = r^*(g\sigma').$$

Moreover, by Corollary 8 we have for the $(n-1)$ -form $g\sigma'$ on S^{n-1} ,

$$\begin{aligned}\int_{S^{n-1}} g\sigma' &= \int_B f dx^1 \wedge \cdots \wedge dx^n \\ &= \int_{\mathbb{R}^n} \omega = 0.\end{aligned}$$

Thus, by the hypothesis for *Step 2*,

$$g\sigma' = d\lambda \quad \text{for some } (n-2)\text{-form } \lambda \text{ on } S^{n-1}.$$

Hence

$$\eta = r^*(d\lambda) = d(r^*\lambda).$$

Let $h: \mathbb{R}^n \rightarrow [0, 1]$ be any C^∞ function with $h = 1$ on A and $h = 0$ in a neighborhood of 0. Then $hr^*\lambda$ is a C^∞ form on \mathbb{R}^n and

$$\omega = d\eta = d(\eta - d(hr^*\lambda)):$$

the form $\eta - d(hr^*\lambda)$ has compact support, since on A we have

$$\eta - d(hr^*\lambda) = \eta - d(r^*\lambda) = 0.$$

Step 3. Choose an n -form ω such that $\int_M \omega \neq 0$ and ω has compact support contained in an open set $U \subset M$, with U diffeomorphic to \mathbb{R}^n . If ω' is any other n -form with compact support, we want to show that there is a number c and a form η with compact support such that

$$\omega' = c\omega + d\eta.$$

Using a partition of unity, we can write

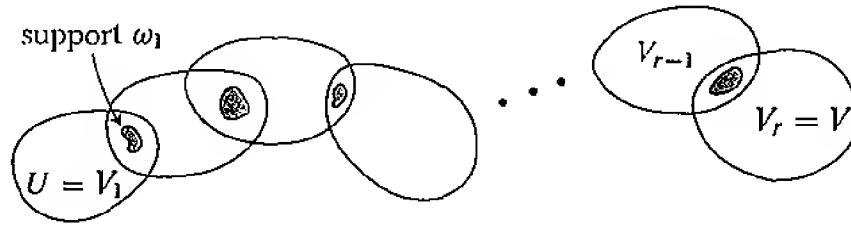
$$\omega' = \phi_1 \omega' + \cdots + \phi_k \omega'$$

where each $\phi_i \omega'$ has compact support contained in some open set $U_i \subset M$ with U_i diffeomorphic to \mathbb{R}^n . It obviously suffices to find c_i and η_i with $\phi_i \omega' = c_i \omega + d\eta_i$, for each i . In other words, we can assume ω' has support contained in some open $V \subset M$ which is diffeomorphic to \mathbb{R}^n .

Using the connectedness of M , it is easy to see that there is a sequence of open sets

$$U = V_1, \dots, V_r = V$$

diffeomorphic to \mathbb{R}^n , with $V_i \cap V_{i+1} \neq \emptyset$. Choose forms ω_i with support $\omega_i \subset$



$V_i \cap V_{i+1}$ and $\int_{V_i} \omega_i \neq 0$. Since we are assuming the theorem for \mathbb{R}^n we have

$$\begin{aligned} \omega_1 - c_1 \omega &= d\eta_1 \\ \omega_2 - c_2 \omega_1 &= d\eta_2 \\ &\vdots \\ \omega' - c_r \omega_{r-1} &= d\eta_r, \end{aligned}$$

where all η_i have compact support ($\subset V_i$). From this we clearly obtain the desired result. ♦

The method used in the last step can be used to derive another result.

10. THEOREM. If M is any connected non-orientable n -manifold, then $H_c^n(M) = 0$.

PROOF. Choose an n -form ω with compact support contained in an open set U diffeomorphic to \mathbb{R}^n , such that $\int_U \omega \neq 0$ (this integral makes sense, since U is orientable). It obviously suffices to show that $\omega = d\eta$ for some form η with compact support. Consider a sequence

$$U = V_1, \dots, V_r = V$$

of coordinate systems (V_i, x_i) where each $x_i \circ x_{i+1}^{-1}$ is orientation preserving. Choose the forms ω_i in *Step 3* so that, using the orientation of V_i which makes $x_i: V_i \rightarrow \mathbb{R}^n$ orientation preserving, we have $\int_{V_i} \omega_i > 0$; then also $\int_{V_{i+1}} \omega_i > 0$. Consequently, the numbers

$$c_i = \int_{V_i} \omega_i \bigg/ \int_{V_i} \omega_{i-1} \quad \text{are positive.}$$

It follows that

$$\omega_i = c\omega + d\eta \quad \text{where } c > 0.$$

Now if M is unorientable, there is such a sequence where $V_r = V_1$ but $x_r \circ x_1^{-1}$ is orientation reversing. Taking $\omega' = -\omega$, we have

$$-\omega = c\omega + d\eta \quad \text{for } c > 0$$

so

$$(-c - 1)\omega = d\eta \quad \text{for } -c - 1 \neq 0. \quad \spadesuit$$

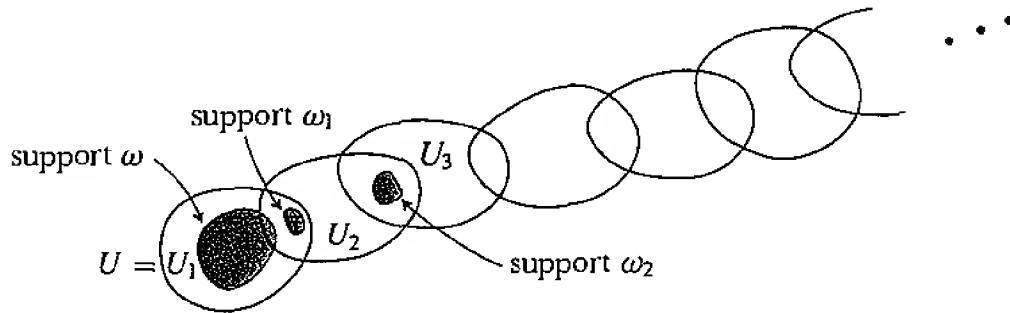
We can also compute $H^n(M)$ for non-compact M .

11. THEOREM. If M is a connected non-compact n -manifold (orientable or not), then $H^n(M) = 0$.

PROOF. Consider first an n -form ω with support contained in a coordinate neighborhood U which is diffeomorphic to \mathbb{R}^n . Since M is not compact, there is an infinite sequence

$$U = U_1, U_2, U_3, U_4, \dots$$

of such coordinate neighborhoods such that $U_i \cap U_{i+1} \neq \emptyset$, and such that the sequence is eventually in the complement of any compact set.



Now choose n -forms ω_i with compact support contained in $U_i \cap U_{i+1}$, such that $\int_{U_i} \omega_i \neq 0$. There are constants c_i and forms η_i with compact support $\subset U_i$ such that

$$\begin{aligned}\omega &= c_1 \omega_1 + d\eta_1 \\ \omega_i &= c_{i+1} \omega_{i+1} + d\eta_{i+1} \quad i \geq 1.\end{aligned}$$

Then

$$\begin{aligned}\omega &= d\eta_1 + c_1 \omega_1 \\ &= d\eta_1 + c_1 d\eta_2 + c_1 c_2 \omega_2 \\ &= d\eta_1 + c_1 d\eta_2 + c_1 c_2 d\eta_3 + c_1 c_2 c_3 \omega_3 \\ &= \dots\end{aligned}$$

Since any point $p \in M$ is eventually in the complement of the U_i 's, we have

$$\omega = d\eta_1 + c_1 d\eta_2 + c_1 c_2 d\eta_3 + c_1 c_2 c_3 d\eta_4 + \dots,$$

where the right side makes sense since the U_i are eventually outside of any compact set.

Now it can be shown (Problem 20) that there is actually such a sequence U_1, U_2, U_3, \dots whose union is all of M (repetitions are allowed, and U_i may intersect several U_j for $j < i$, but the sequence is still eventually outside of any compact set). The cover $\mathcal{O} = \{U_i\}$ is then locally finite. Let $\{\phi_{U_i}\}$ be a partition of unity subordinate to \mathcal{O} . If ω is an n -form on M , then for each U_i we have seen that

$$\phi_{U_i} \omega = d\eta_i \quad \text{where } \eta_i \text{ has support contained in } U_i \cup U_{i+1} \cup U_{i+2} \cup \dots.$$

Hence

$$\omega = \sum_{i=1}^{\infty} \phi_{U_i} \omega = \sum_{i=1}^{\infty} d\eta_i = d \left(\sum_{i=1}^{\infty} \eta_i \right). \quad \spadesuit$$

SUMMARY OF RESULTS

(1) For \mathbb{R}^n we have

$$H^k(\mathbb{R}^n) \approx \begin{cases} \mathbb{R} & k = 0 \\ 0 & k > 0. \end{cases}$$

(2) If M is a connected n -manifold, then

$$\begin{aligned} H^0(M) &\approx \mathbb{R} \\ H_c^n(M) &\approx \begin{cases} \mathbb{R} & \text{if } M \text{ is orientable} \\ 0 & \text{if } M \text{ is non-orientable} \end{cases} \\ H^n(M) &\approx \begin{cases} H_c^n & \text{if } M \text{ is compact} \\ 0 & \text{if } M \text{ is not compact.} \end{cases} \end{aligned}$$

We also know that $H^{n-1}(\mathbb{R}^n - \{0\}) \neq 0$, but we have not listed this result, since we will eventually improve it. In order to proceed further with our computations we need to examine the behavior of the de Rham cohomology vector spaces under C^∞ maps $f: M \rightarrow N$. If ω is a closed k -form on N , then $f^*\omega$ is also closed ($df^*\omega = f^*d\omega = 0$), so f^* takes $Z^k(N)$ to $Z^k(M)$. On the other hand, f^* also takes $B^k(N)$ to $B^k(M)$, since $f^*(d\eta) = d(f^*\eta)$. This shows that f^* induces a map

$$Z^k(N)/B^k(N) \rightarrow Z^k(M)/B^k(M),$$

also denoted by f^* :

$$f^*: H^k(N) \rightarrow H^k(M).$$

For example, consider the case $k = 0$. If N is connected, then $H^0(N)$ is just the collection of constant functions $c: N \rightarrow \mathbb{R}$. Then $f^*(c) = c \circ f$ is also a constant function. If M is connected, then $f^*: H^0(N) \rightarrow H^0(M)$ is just the identity map under the natural identification of $H^0(N)$ and $H^0(M)$ with \mathbb{R} . If M is disconnected, with components M_α , $\alpha \in A$, then $H^0(M)$ is isomorphic to the direct sum

$$\bigoplus_{\alpha \in A} \mathbb{R}_\alpha, \quad \text{where each } \mathbb{R}_\alpha \approx \mathbb{R};$$

the map f^* takes $c \in \mathbb{R}$ into the element of $\bigoplus \mathbb{R}_\alpha$ with α^{th} component equal to c . If N is also disconnected, with components N_β , $\beta \in B$, then

$$f^*: \bigoplus_{\beta \in B} \mathbb{R}_\beta \rightarrow \bigoplus_{\alpha \in A} \mathbb{R}_\alpha$$

takes the element $\{c_\beta\}$ of $\bigoplus_{\beta \in B} \mathbb{R}_\beta$ to $\{c'_\alpha\}$, where $c'_\alpha = c_\beta$ when $f(M_\alpha) \subset N_\beta$.

A more interesting case, and the only one we are presently in a position to look at, is the map

$$f^*: H^n(N) \rightarrow H^n(M)$$

when M and N are both compact connected oriented n -manifolds. There is no natural way to make $H^n(M)$ isomorphic to \mathbb{R} , so we really want to compare

$$\int_M f^* \omega \quad \text{and} \quad \int_N \omega$$

for ω an n -form on N . Choose one ω_0 with $\int_N \omega_0 \neq 0$. Then there is some number a such that

$$\int_M f^* \omega_0 = a \cdot \int_N \omega_0.$$

Since $\omega \mapsto \int_M \omega$ is an isomorphism of $H^n(M)$ and \mathbb{R} (and similarly for N) it follows that for *every* form ω we have

$$\int_M f^* \omega = a \cdot \int_N \omega.$$

The number $a = \deg f$, which depends only on f , is called the **degree** of f . If M and N are not compact, but f is proper (the inverse image of any compact set is compact), then we have a map

$$f^*: H_c^n(N) \rightarrow H_c^n(M)$$

and a number $\deg f$, such that

$$\int_M f^* \omega = (\deg f) \int_N \omega$$

for all forms ω on N with compact support. Until one sees the proof of the next theorem, it is almost unbelievable that this number is *always an integer*.

12. THEOREM. Let $f: M \rightarrow N$ be a proper map between two connected oriented n -manifolds (M, μ) and (N, ν) . Let $q \in N$ be a regular value of f . For each $p \in f^{-1}(q)$, let

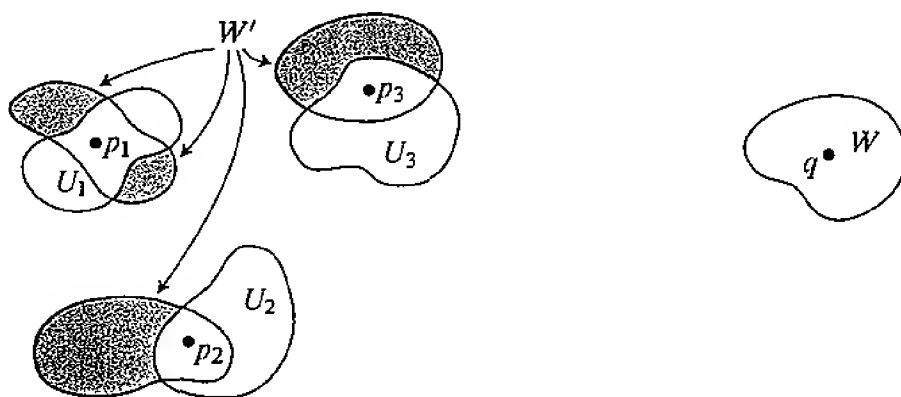
$$\text{sign}_p f = \begin{cases} 1 & \text{if } f_{*p}: M_p \rightarrow N_q \text{ is orientation preserving} \\ & \text{(using the orientations } \mu_p \text{ for } M_p \text{ and } \nu_q \text{ for } N_q) \\ -1 & \text{if } f_{*p} \text{ is orientation reversing.} \end{cases}$$

Then

$$\deg f = \sum_{p \in f^{-1}(q)} \text{sign}_p f \quad (= 0 \text{ if } f^{-1}(q) = \emptyset).$$

PROOF. Notice first that regular values exist, by Sard's Theorem. Moreover, $f^{-1}(q)$ is finite, since it is compact and consists of isolated points, so the sum above is a finite sum.

Let $f^{-1}(q) = \{p_1, \dots, p_k\}$. Choose coordinate systems (U_i, x_i) around p_i such that all points in U_i are regular values of f , and the U_i are disjoint. We want to choose a coordinate system (V, y) around q such that $f^{-1}(V) = U_1 \cup \dots \cup U_k$. To do this, first choose a compact neighborhood W of q , and let



$W' \subset M$ be the compact set

$$W' = f^{-1}(W) - (U_1 \cup \dots \cup U_k).$$

Then $f(W')$ is a closed set which does not contain q . We can therefore choose $V \subset W - f(W')$. This ensures that $f^{-1}(V) \subset U_1 \cup \dots \cup U_k$. Finally, redefine U_i to be $U_i \cap f^{-1}(V)$.

Now choose ω on N to be $\omega = g dy^1 \wedge \dots \wedge dy^n$ where $g \geq 0$ has compact support contained in V . Then support $f^*\omega \subset U_1 \cup \dots \cup U_k$. So

$$\int_M f^*\omega = \sum_{i=1}^k \int_{U_i} f^*\omega.$$

Since f is a diffeomorphism from each U_i to V we have

$$\begin{aligned} \int_{U_i} f^*\omega &= \int_V \omega \quad \text{if } f \text{ is orientation preserving} \\ &= - \int_V \omega \quad \text{if } f \text{ is orientation reversing.} \end{aligned}$$

Since f is orientation preserving [or reversing] precisely when $\text{sign}_p f = 1$ [or -1] this proves the theorem. ♦

As an immediate application of the theorem, we compute the degree of the “antipodal map” $A: S^n \rightarrow S^n$ defined by $A(p) = -p$. We have already seen that A is orientation preserving or reversing at all points, depending on whether n is odd or even. Since $A^{-1}(p)$ consists of just one point, we conclude that

$$\deg A = (-1)^{n-1}.$$

We can draw an interesting conclusion from this result, but we need to introduce another important concept first. Two functions $f, g: M \rightarrow N$ between two C^∞ manifolds are called (smoothly) homotopic if there is a smooth function

$$H: M \times [0, 1] \rightarrow N$$

with

$$\begin{aligned} H(p, 0) &= f(p) \\ H(p, 1) &= g(p) \end{aligned} \quad \text{for all } p \in M;$$

the map H is called a (smooth) homotopy between f and g . Notice that M is smoothly contractible to a point $p_0 \in M$ if and only if the identity map of M is homotopic to the constant map p_0 . Recall that for every k -form ω on $M \times [0, 1]$ we defined a $(k-1)$ -form $I\omega$ on M such that

$$i_1^* \omega - i_0^* \omega = d(I\omega) + I(d\omega).$$

We used this fact to show that all closed forms on a smoothly contractible manifold are exact. We can now prove a more general result.

13. THEOREM. If $f, g: M \rightarrow N$ are smoothly homotopic, then the maps

$$f^*: H^k(N) \rightarrow H^k(M)$$

$$g^*: H^k(N) \rightarrow H^k(M)$$

are equal, $f^* = g^*$.

PROOF. By assumption, there is a smooth map $H: M \times [0, 1] \rightarrow N$ with

$$f = H \circ i_0$$

$$g = H \circ i_1.$$

Any element of $H^k(N)$ is the equivalence class $[\omega]$ of some closed k -form ω on N . Then

$$\begin{aligned} g^* \omega - f^* \omega &= (H \circ i_1)^* \omega - (H \circ i_0)^* \omega \\ &= i_1^*(H^* \omega) - i_0^*(H^* \omega) \\ &= d(IH^* \omega) + I(dH^* \omega) \\ &= d(IH^* \omega) + 0. \end{aligned}$$

But this means that $g^*([\omega]) = f^*([\omega])$. ♦

14. COROLLARY. If M and N are compact oriented n -manifolds and the maps $f, g: M \rightarrow N$ are homotopic, then $\deg f = \deg g$.

15. COROLLARY. If n is even, then there does not exist a nowhere zero vector field on S^n .

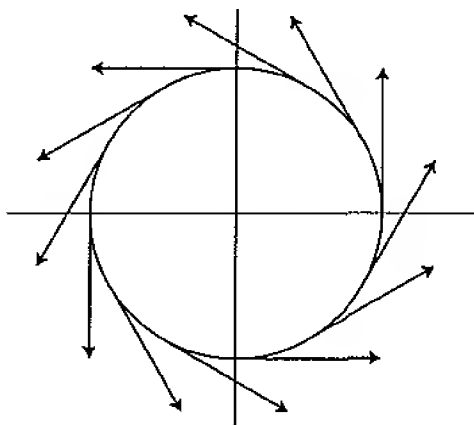
PROOF. We have already seen that the degree of the antipodal map $A: S^n \rightarrow S^n$ is $(-1)^{n-1}$. Since the identity map has degree 1, A is not homotopic to the identity for n even. But if there is a nowhere zero vector field on S^n , then we can construct a homotopy between A and the identity map as follows. For each p , there is a unique great semi-circle γ_p from p to $A(p) = -p$ whose tangent vector at p is a multiple of $X(p)$. Define

$$H(p, t) = \gamma_p(t). \quad \spadesuit$$

For n odd we can explicitly construct a nowhere zero vector field on S^n . For $p = (x_1, \dots, x_{n+1}) \in S^n$ we define

$$X(p) = (-x_1, x_2, -x_3, x_4, \dots, -x_{n+1}, x_n);$$

this is perpendicular to $p = (x_1, x_2, \dots, x_{n+1})$, and therefore in S^n_p . (On S^1 this gives the standard picture.) The vector field on S^n can then be used to give



a homotopy between A and the identity map.

For another application of Theorem 13, consider the retraction

$$r: \mathbb{R}^n - \{0\} \rightarrow S^{n-1} \quad r(p) = p/|p|.$$

If $i: S^{n-1} \rightarrow \mathbb{R}^n - \{0\}$ is the inclusion, then

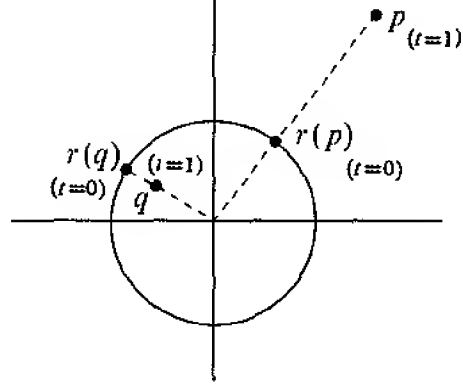
$$r \circ i: S^{n-1} \rightarrow S^{n-1} \text{ is the identity } 1 \text{ of } S^{n-1}.$$

The map

$$i \circ r: \mathbb{R}^n - \{0\} \rightarrow \mathbb{R}^n - \{0\} \quad i \circ r(p) = p/|p|$$

is, of course, not the identity, but it is homotopic to the identity; we can define the homotopy H by

$$H(p, t) = tp + (1 - t)r(p) \in \mathbb{R}^n - \{0\}.$$



A retraction with this property is called a **deformation retraction**. Whenever r is a deformation retraction, the maps $(r \circ i)^*$ and $(i \circ r)^*$ are the identity. Thus, for the case of $S^{n-1} \subset \mathbb{R}^n - \{0\}$, we have

$$\begin{aligned} H^k(S^{n-1}) &\xrightarrow{r^*} H^k(\mathbb{R}^n - \{0\}) \\ H^k(\mathbb{R}^n - \{0\}) &\xrightarrow{i^*} H^k(S^{n-1}) \end{aligned}$$

and

$$\begin{aligned} r^* \circ i^* &= (i \circ r)^* = \text{identity of } H^k(\mathbb{R}^n - \{0\}) \\ i^* \circ r^* &= (r \circ i)^* = \text{identity of } H^k(S^{n-1}). \end{aligned}$$

So i^* and r^* are inverses of each other. Thus

$$H^k(S^{n-1}) \approx H^k(\mathbb{R}^n - \{0\}) \quad \text{for all } k.$$

In particular, we have $H^{n-1}(\mathbb{R}^n - \{0\}) \approx \mathbb{R}$. A generator of $H^{n-1}(\mathbb{R}^n - \{0\})$ is the closed form $r^*\sigma'$.

We are now going to compute $H^k(\mathbb{R}^n - \{0\})$ for all k . We need one further observation. The manifold

$$M \times \{0\} \subset M \times \mathbb{R}^l$$

is clearly a deformation retraction of $M \times \mathbb{R}^l$. So $H^k(M) \approx H^k(M \times \mathbb{R}^l)$ for all l .

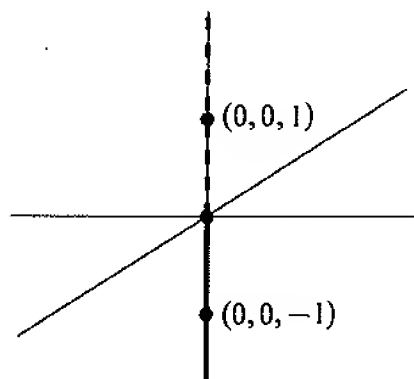
16. THEOREM. For $0 < k < n - 1$ we have $H^k(\mathbb{R}^n - \{0\}) = H^k(S^{n-1}) = 0$.

PROOF. Induction on n . The first case where there is anything to prove is $n = 3$. We claim $H^1(\mathbb{R}^3 - \{0\}) = 0$.

Let ω be a closed 1-form on \mathbb{R}^3 . Let A and B be the open sets

$$A = \mathbb{R}^3 - \{(0, 0) \times (-\infty, 0]\}$$

$$B = \mathbb{R}^3 - \{(0, 0) \times [0, \infty)\}.$$



Since A and B are both star-shaped (with respect to the points $(0, 0, 1)$ and $(0, 0, -1)$, respectively), there are 0-forms f_A and f_B on A and B with

$$\omega = df_A \quad \text{on } A$$

$$\omega = df_B \quad \text{on } B.$$

Now

$$d(f_A - f_B) = 0 \quad \text{on } A \cap B,$$

and

$$A \cap B = [\mathbb{R}^2 - \{0\}] \times \mathbb{R},$$

so clearly $f_A - f_B$ is a constant c on $A \cap B$. Thus ω is exact, for

$$\omega = d(f_A - c) \quad \text{on } A$$

$$\omega = d(f_B) \quad \text{on } B$$

and $f_A - c = f_B$ on $A \cap B$.

If ω is a closed 1-form on \mathbb{R}^4 , there is a similar argument, using

$$A = \mathbb{R}^4 - \{(0, 0, 0) \times (-\infty, 0]\}$$

$$B = \mathbb{R}^4 - \{(0, 0, 0) \times [0, \infty)\}.$$

If ω is a closed 2-form on \mathbb{R}^4 , then we obtain 1-forms η_A and η_B with

$$\omega = d\eta_A \quad \text{on } A$$

$$\omega = d\eta_B \quad \text{on } B.$$

Now

$$d(\eta_A - \eta_B) = 0 \quad \text{on } A \cap B$$

and

$$H^1(A \cap B) = H^1([\mathbb{R}^3 - \{0\}] \times \mathbb{R}) \approx H^1(\mathbb{R}^3 - \{0\}) = 0.$$

So $\eta_A - \eta_B = d\lambda$ for some 0-form λ on $A \cap B$. Unlike the previous case, we cannot simply consider $\eta_A - d\lambda$, since this is not defined on A . To circumvent this difficulty, note that there is a partition of unity $\{\phi_A, \phi_B\}$ for the cover $\{A, B\}$ of $\mathbb{R}^3 - \{0\}$:

$$\begin{aligned} \phi_A + \phi_B &= 1 \\ d\phi_A + d\phi_B &= 0 \\ \text{support } \phi_A &\subset A \\ \text{support } \phi_B &\subset B. \end{aligned}$$

Now, if

$$\phi_B \lambda \quad \text{denotes} \quad \begin{cases} \phi_B \lambda & \text{on } A \cap B \\ 0 & \text{on } A - (A \cap B), \end{cases}$$

and similarly for $\phi_A \lambda$, then

$$\begin{aligned} \phi_B \lambda &\text{ is a } C^\infty \text{ form on } A \\ \phi_A \lambda &\text{ is a } C^\infty \text{ form on } B. \end{aligned}$$

On $A \cap B$ we have

$$\begin{aligned} \eta_A - d(\phi_B \lambda) &= \eta_A - \phi_B d\lambda - d\phi_B \wedge \lambda \\ &= \eta_A + (\phi_A - 1) d\lambda + d\phi_A \wedge \lambda \\ &= \eta_A - d\lambda + d(\phi_A \lambda) \\ &= \eta_B + d(\phi_A \lambda). \end{aligned}$$

So we can define a C^∞ form on $\mathbb{R}^n - \{0\} = A \cup B$ by letting it be $\eta_A - d(\phi_B \lambda)$ on A , and $\eta_B + d(\phi_A \lambda)$ on B . Clearly,

$$\begin{aligned} \omega &= d\eta_A = d(\eta_A - d(\phi_B \lambda)) \quad \text{on } A \\ &= d\eta_B = d(\eta_B + d(\phi_A \lambda)) \quad \text{on } B, \end{aligned}$$

so ω is exact.

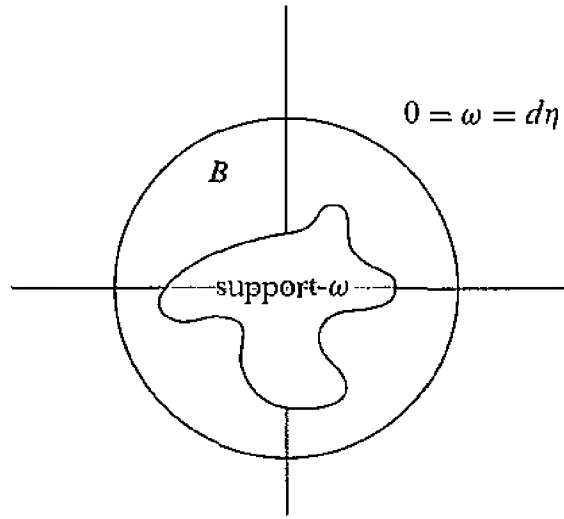
The general inductive step is similar. ♦

We end this chapter with one more calculation, which we will need in Chapter 11.

17. THEOREM. For $0 \leq k < n$ we have $H_c^k(\mathbb{R}^n) = 0$.

PROOF. The proof that $H_c^0(\mathbb{R}^n) = 0$ is left to the reader.

Let ω be a k -form on \mathbb{R}^n with compact support, $0 < k < n$. We know that $\omega = d\eta$ for some $(k-1)$ -form η on \mathbb{R}^n . Let B be a closed ball containing support ω . Then on $A = \mathbb{R}^n - B$ we have $d\eta = 0$. Since A is diffeomorphic to



$\mathbb{R}^n - \{0\}$ and $k-1 < n-1$ we have from Theorem 16 that

$$\eta = d\lambda \quad \text{for some } (k-2)\text{-form } \lambda \text{ on } A.$$

Let $f: \mathbb{R}^n \rightarrow [0, 1]$ be a C^∞ function with $f = 0$ in a neighborhood of B and $f = 1$ on $\mathbb{R}^n - 2B$, where $2B$ denotes the ball of twice the radius of B . Then $d(f\lambda)$ makes sense on all of \mathbb{R}^n and

$$\omega = d\eta = d(\eta - d(f\lambda));$$

the form $\eta - d(f\lambda)$ clearly has compact support contained in $2B$. ♦

PROBLEMS

1. *The Riemann integral versus the Darboux integral.* Let $f : [a, b] \rightarrow \mathbb{R}$ be bounded. For a partition $P = \{t_0 < \cdots < t_n\}$ of $[a, b]$, let $m_i = m_i(f)$ be the inf of f on $[t_{i-1}, t_i]$ and define $M_i = M_i(f)$ similarly. A *choice* for P is an n -tuple $\xi = (\xi_1, \dots, \xi_n)$ with $\xi_i \in [t_{i-1}, t_i]$. We define the “lower sum”, “upper sum”, and “Riemann sum” for a partition P and choice ξ by

$$L(f, P) = \sum_{i=1}^n m_i(f) \cdot (t_i - t_{i-1})$$

$$U(f, P) = \sum_{i=1}^n M_i(f) \cdot (t_i - t_{i-1})$$

$$S(f, P, \xi) = \sum_{i=1}^n f(\xi_i)(t_i - t_{i-1}).$$

Clearly $L(f, P) \leq S(f, P, \xi) \leq U(f, P)$. We call f **Darboux integrable** if the sup of all $L(f, P)$ equals the inf of all $U(f, P)$; this sup or inf is called the **Darboux integral** of f on $[a, b]$. We call f **Riemann integrable** if

$$\lim_{\|P\| \rightarrow 0} S(f, P, \xi) \text{ exists;}$$

the limit is called the **Riemann integral** of f on $[a, b]$.

(a) We can define $S(f, P, \xi)$ even if f is not bounded. Show however, that $\lim_{\|P\| \rightarrow 0} S(f, P, \xi)$ cannot exist if f is unbounded.

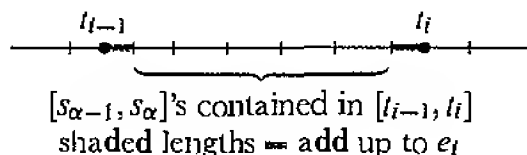
(b) If f is continuous on $[a, b]$, then f is Riemann and Darboux integrable on $[a, b]$, and the two integrals are equal. (Use uniform continuity of f on $[a, b]$.)

(c) If f is Riemann integrable on $[a, b]$, then f is Darboux integrable on $[a, b]$ and the two integrals are equal.

(d) Let $m \leq f \leq M$ on $[a, b]$. Let $P = \{s_0 < \cdots < s_m\}$ and $Q = \{t_0 < \cdots < t_n\}$ be two partitions of $[a, b]$. For each $i = 1, \dots, n$, let

$e_i = \text{length of } [t_{i-1}, t_i]$

– sum of lengths of all $[s_{\alpha-1}, s_\alpha]$ which are contained in $[t_{i-1}, t_i]$.



Show that, if M_i denotes the sup of f on $[t_{i-1}, t_i]$, then

$$\begin{aligned} U(f, P) &\leq U(f, Q) + \sum_{i=1}^n (M - M_i) e_i \\ &\leq U(f, Q) + (M - m) \sum_{i=1}^n e_i. \end{aligned}$$

There is a similar result for lower sums.

(e) Show that $\sum_{i=1}^n e_i \rightarrow 0$ as $\|P\| \rightarrow 0$, and deduce Darboux's Theorem:

$$\begin{aligned} \lim_{\|P\| \rightarrow 0} U(f, P) &= \inf \{U(f, Q) : Q \text{ a partition of } [a, b]\} \\ \lim_{\|P\| \rightarrow 0} L(f, P) &= \sup \{L(f, Q) : Q \text{ a partition of } [a, b]\}. \end{aligned}$$

(f) If f is Darboux integrable on $[a, b]$, then f is Riemann integrable on $[a, b]$.

(g) (Osgood's Theorem). Let f and g be integrable on $[a, b]$. Show that for choices ξ, ξ' for P ,

$$\lim_{\|P\| \rightarrow 0} \sum_{i=1}^n f(\xi_i) g(\xi'_i) (t_i - t_{i-1}) = \int_a^b f g.$$

Hint: If $|g| \leq M$ on $[a, b]$, then $|f(\xi'_i)g(\xi'_i) - f(\xi_i)g(\xi'_i)| \leq M|f(\xi'_i) - f(\xi_i)|$.

(h) Show that $\int_c f dx + g dy$, defined as a limit of sums, equals

$$\int_a^b [f(c(t))c^{1'}(t) + g(c(t))c^{2'}(t)] dt.$$

2. Compute $\int_c d\theta = \int_{[0,1]} c^* d\theta$, where $c(t) = (\cos 2\pi t, \sin 2\pi t)$ on $[0, 1]$.

3. For n an integer, and $R > 0$, let $c_{R,n} : [0, 1] \rightarrow \mathbb{R}^2 - \{0\}$ be defined by

$$c_{R,n}(t) = (R \cos 2n\pi t, R \sin 2n\pi t).$$

(a) Show that there is a singular 2-cube $c : [0, 1]^2 \rightarrow \mathbb{R}^2 - \{0\}$ such that $c_{R_1,n} - c_{R_2,n} = \partial c$.

(b) If $c : [0, 1] \rightarrow \mathbb{R}^2 - \{0\}$ is any curve with $c(0) = c(1)$, show that there is some n such that $c - c_{1,n}$ is a boundary in $\mathbb{R}^2 - \{0\}$.

(c) Show that n is unique. It is called the **winding number** of c around 0.

4. Let $f: \mathbb{C} \rightarrow \mathbb{C}$ be a polynomial, $f(z) = z^n + a_1 z^{n-1} + \cdots + a_n$, where $n \geq 1$. Define $c_{R,f}: [0, 1] \rightarrow \mathbb{C}$ by $c_{R,f} = f \circ c_{R,1}$.

(a) Show that if R is large enough, then $c_{R,f} - c_{R,n}$ is the boundary of a chain in $\mathbb{C} - \{0\}$. *Hint:* Note that $c_{R^n,n}(t) = [c_{R,1}(t)]^n$, and write

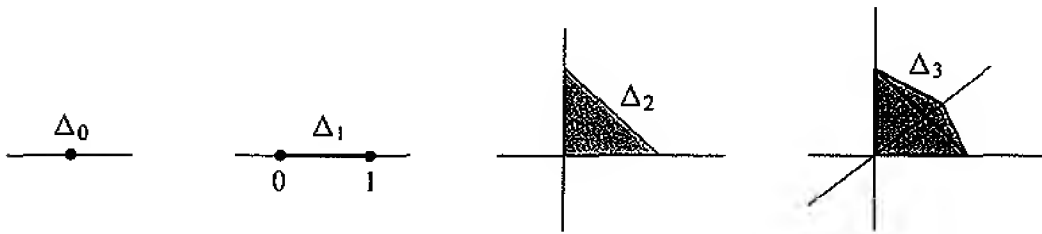
$$f(z) = z^n \left(1 + \frac{a_1}{z} + \cdots + \frac{a_n}{z^n} \right).$$

(b) Show that $f(z) = 0$ for some $z \in \mathbb{C}$ ("Fundamental Theorem of Algebra"). *Hint:* If $f(z) \neq 0$ for all z with $|z| \leq R$, then $c_{R,f} - c_{0,f}$ is a boundary.

5. Some approaches to integration use singular simplexes instead of singular cubes. Although Stokes' Theorem becomes more complicated, there are some advantages in using singular simplexes, as indicated in the next Problem.

Let $\Delta_n \subset \mathbb{R}^n$ be the set of all $x \in \mathbb{R}^n$ such that

$$0 \leq x^i \leq 1, \quad \sum_{i=1}^n x^i \leq 1.$$



A singular n -simplex in M is a C^∞ function $c: \Delta_n \rightarrow M$, and an n -chain is a formal sum of singular n -simplexes. As before, let $I^n: \Delta_n \rightarrow \mathbb{R}^n$ be the inclusion map. Define $\partial_i: \Delta_{n-1} \rightarrow \Delta_n$ by

$$\begin{aligned} \partial_0(x) &= ([1 - \sum_{i=1}^{n-1} x^i], x^1, \dots, x^{n-1}) \\ \partial_i(x) &= (x^1, \dots, x^{i-1}, 0, x^i, \dots, x^{n-1}) \quad 0 < i \leq n, \end{aligned}$$

and for singular n -simplexes c , define $\partial_i c = c \circ \partial_i$. Then we define

$$\partial c = \sum_{i=0}^n (-1)^i \partial_i c.$$

- Describe geometrically the images $\partial_i(\Delta_{n-1})$ in Δ_n .
- Show that $\partial^2 = 0$.

(c) Show that if $\omega = f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n$ is an $(n-1)$ -form on \mathbb{R}^n , then

$$\int_{I^n} d\omega = \int_{\partial I^n} \omega.$$

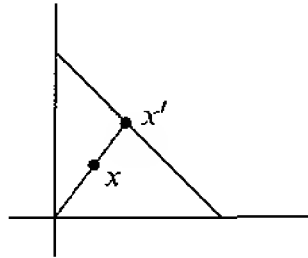
(Imitate the proof for cubes.)

(d) Define $\int_c \omega$ for any k -chain c in M and k -form ω on M , and prove that

$$\int_c d\omega = \int_{\partial c} \omega$$

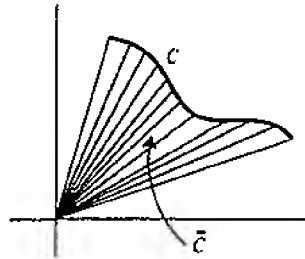
for any $(k-1)$ -form ω .

6. Every $x \in \Delta_{k+1}$ can be written as tx' , for $0 \leq t \leq 1$, and $x' \in \partial_0(\Delta_k)$.



Moreover, x' is unique except when $t = 0$. For any singular k -simplex $c: \Delta_k \rightarrow \mathbb{R}^n$, define $\bar{c}: \Delta_{k+1} \rightarrow \mathbb{R}^n$ by

$$\bar{c}(x) = t \cdot c(x').$$



We then define \bar{c} for chains c in the obvious way.

(a) Show that $\partial c = 0$ implies that $c = \partial \bar{c}$.

(b) Let $c: [0, 1] \rightarrow \mathbb{R}^2$ be a closed curve. Show that c is *not* the boundary of any sum σ of singular 2-cubes. *Hint:* If $\partial \sigma = \sum_i a_i c_i$, what can be said about $\sum_i a_i$?

(c) Show that we do have $c = \partial \sigma + c'$ where c' is *degenerate*, that is, $c'([0, 1])$ is a point.

(d) If $c_1(0) = c_2(0)$ and $c_1(1) = c_2(1)$, show that $c_1 - c_2$ is a boundary, using either simplexes or cubes.

7. Let ω be a 1-form on a manifold M . Suppose that $\int_c \omega = 0$ for every closed curve c in M . Show that ω is exact. *Hint:* If we do have $\omega = df$, then for any curve c we have

$$\int_c \omega = f(c(1)) - f(c(0)).$$

8. A manifold M is called **simply-connected** if M is connected and if every smooth map $f: S^1 \rightarrow M$ is smoothly contractible to a point. [Actually, any space M (not necessarily a manifold) is called simply-connected if it is connected and any continuous $f: S^1 \rightarrow M$ is (continuously) contractible to a point. It is not hard to show that for a manifold we may insert “smooth” at both places.]

(a) If M is smoothly contractible to a point, then M is simply-connected.

(b) S^1 is not simply-connected.

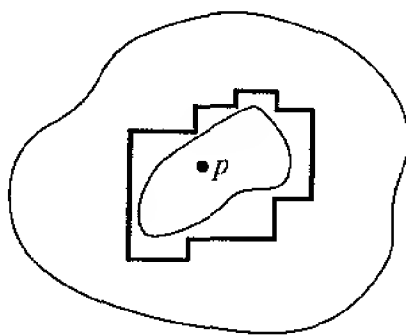
(c) S^n is simply-connected for $n > 1$. *Hint:* Show that a smooth $f: S^1 \rightarrow S^n$ is not onto.

(d) If M is simply-connected and $p \in M$, then any smooth map $f: S^1 \rightarrow M$ is smoothly contractible to p .

(e) If $M = U \cup V$ where U and V are simply-connected open subsets with $U \cap V$ connected, then M is simply-connected. (This gives another proof that S^n is simply-connected for $n > 1$.) *Hint:* Given $f: S^1 \rightarrow M$, partition S^1 into a finite number of intervals each of which is taken into either U or V .

(f) If M is simply-connected, then $H^1(M) = 0$. (See Problem 7.)

9. (a) Let $U \subset \mathbb{R}^2$ be a bounded open set such that $\mathbb{R}^2 - U$ is not connected. Show that U is not smoothly contractible to a point. (Converse of



Problem 7-24.) *Hint:* If p is in a bounded component of $\mathbb{R}^2 - U$, show that there is a curve in U which “surrounds” p .

(b) A bounded connected open set $U \subset \mathbb{R}^2$ is smoothly contractible to a point if and only if it is simply-connected.

(c) This is false for open subsets of \mathbb{R}^3 .

10. Let ω be an n -form on an oriented manifold M^n . Let Φ and Ψ be two partitions of unity by functions with compact support, and suppose that

$$\sum_{\phi \in \Phi} \int_M \phi \cdot |\omega| < \infty.$$

(a) This implies that $\sum_{\phi \in \Phi} \int_M \phi \cdot \omega$ converges absolutely.

(b) Show that

$$\sum_{\phi \in \Phi} \int_M \phi \cdot \omega = \sum_{\phi \in \Phi} \sum_{\psi \in \Psi} \int_M \psi \cdot \phi \cdot \omega,$$

and show the same result with ω replaced by $|\omega|$. (Note that for each ϕ , there are only finitely many ψ which are non-zero on support ϕ .)

(c) Show that $\sum_{\psi \in \Psi} \int_M \psi \cdot |\omega| < \infty$, and that

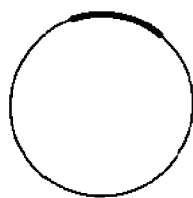
$$\sum_{\phi \in \Phi} \int_M \phi \cdot \omega = \sum_{\psi \in \Psi} \int_M \psi \cdot \omega.$$

We define this common sum to be $\int_M \omega$.

(d) Let $A_n \subset (n, n+1)$ be closed sets. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a C^∞ function with $\int_{A_n} f = (-1)^n/n$ and support $f \subset \bigcup_n A_n$. Find two partitions of unity Φ and Ψ such that $\sum_{\phi \in \Phi} \int_{\mathbb{R}} \phi \cdot f dx$ and $\sum_{\psi \in \Psi} \int_{\mathbb{R}} \psi \cdot f dx$ converge absolutely to different values.

11. Following Problem 7-12, define geometric objects corresponding to odd relative tensors of type $\binom{k}{l}$ and weight w (w any real number).

12. (a) Let M be $\{(x, y) \in \mathbb{R}^2 : |(x, y)| < 1\}$, together with a proper portion



of its boundary, and let $\omega = x dy$. Show that

$$\int_M d\omega \neq \int_{\partial M} \omega,$$

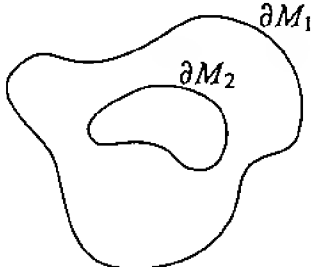
even though both sides make sense, using Problem 10. (No computations needed—note that equality would hold if we had the entire boundary.)

(b) Similarly, find a counterexample to Stokes' Theorem when $M = (0, 1)$ and ω is a 0-form whose support is not compact.

(c) Examine a partition of unity for $(0, 1)$ by functions with compact support to see just why the proof of Stokes' Theorem breaks down in this case.

13. Suppose M is a compact orientable n -manifold (with no boundary), and θ is an $(n-1)$ -form on M . Show that $d\theta$ is 0 at some point.

14. Let $M_1, M_2 \subset \mathbb{R}^n$ be compact n -dimensional manifolds-with-boundary with $M_2 \subset M_1 - \partial M_1$. Show that for any closed $(n-1)$ -form ω on M_1 ,

$$\int_{\partial M_1} \omega = \int_{\partial M_2} \omega.$$


15. Account for the factor $1/|p|^n$ in Lemma 7 (we have $r_*(v_p) = (1/|p|)v_{r(p)}$, but this only accounts for a factor of $1/|p|^{n-1}$, since there are $n-1$ vectors v_1, \dots, v_{n-1}).

16. Use the formula for r^*dx^i (Problem 4-1) to compute $r^*\sigma'$. (Note that

$$r^*\sigma' = r^*i^*\sigma = (i \circ r)^*\sigma;$$

the map $i \circ r: \mathbb{R}^n - \{0\} \rightarrow \mathbb{R}^n - \{0\}$ is just r , considered as a map into $\mathbb{R}^n - \{0\}$.)

17. (a) Let M^n and N^m be oriented manifolds, and let ω and η be an n -form and an m -form with compact support, on M and N , respectively. We will orient $M \times N$ by agreeing that $v_1, \dots, v_n, w_1, \dots, w_m$ is positively oriented in $(M \times N)_{(p,q)} \approx M_p \oplus N_q$ if v_1, \dots, v_n and w_1, \dots, w_m are positively oriented in M_p and N_q , respectively. If $\pi_i: M \times N \rightarrow M$ or N is projection on the i^{th} factor, show that

$$\int_{M \times N} \pi_1^* \omega \wedge \pi_2^* \eta = \int_M \omega \cdot \int_N \eta.$$

(b) If $h: M \times N \rightarrow \mathbb{R}$ is C^∞ , then

$$\int_{M \times N} h \pi_1^* \omega \wedge \pi_2^* \eta = \int_M g \omega,$$

where

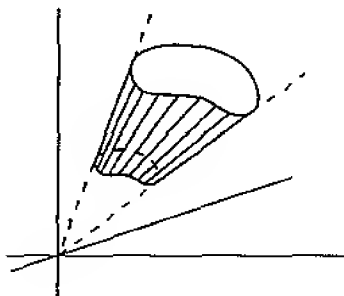
$$g(p) = \int_N h(p, \cdot) \eta, \quad h(p, \cdot) = q \mapsto h(p, q).$$

(c) Every $(m+n)$ -form on $M \times N$ is $h \pi_1^* \omega \wedge \pi_2^* \eta$ for some ω and η .

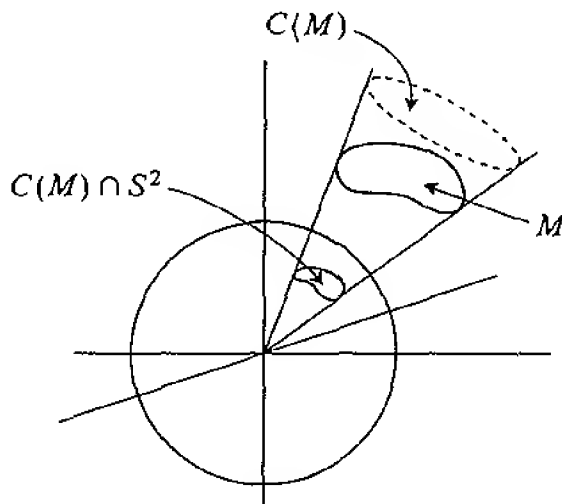
18. (a) Let $p \in \mathbb{R}^n - \{0\}$. Let $w_1, \dots, w_{n-2} \in \mathbb{R}^n_p$ and let $v \in \mathbb{R}^n_p$ be $(\lambda p)_p$ for some $\lambda \in \mathbb{R}$. Show that

$$r^* \sigma'(v, w_1, \dots, w_{n-2}) = 0.$$

(b) Let $M \subset \mathbb{R}^n - \{0\}$ be a compact $(n-1)$ -manifold-with-boundary which is the union of segments of rays through 0. Show that $\int_M r^* \sigma' = 0$.



(c) Let $M \subset \mathbb{R}^n - \{0\}$ be a compact $(n-1)$ -manifold-with-boundary which intersects every ray through 0 at most once, and let $C(M) = \{\lambda p : p \in M, \lambda \geq 0\}$.

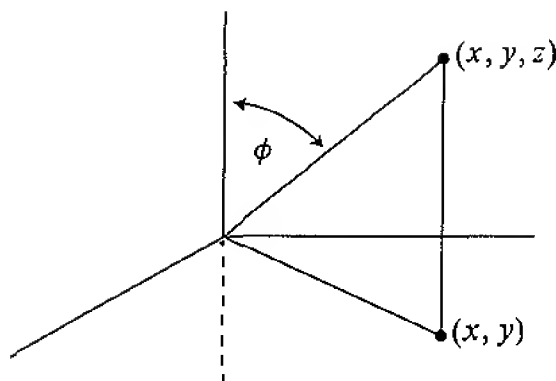


Show that

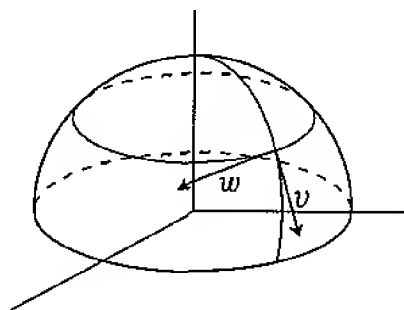
$$\int_M r^* \sigma' = \int_{C(M) \cap S^2} r^* \sigma'.$$

The latter integral is the measure of the solid angle subtended by M . For this reason we often denote $r^* \sigma'$ by $d\Theta_n$.

19. For all $(x, y, z) \in \mathbb{R}^3$ except those with $x = 0, y = 0, z \in (-\infty, 0]$, we define $\phi(x, y, z)$ to be the angle between the positive z -axis and the ray from 0 through (x, y, z) .



- (a) $\phi(x, y, z) = \arctan(\sqrt{x^2 + y^2}/z)$ (with appropriate conventions).
- (b) If $v(p) = |p|$, and θ is considered as a function on \mathbb{R}^3 , $\theta(x, y, z) = \arctan y/x$, then (v, θ, ϕ) is a coordinate system on the set of all points (x, y, z) in \mathbb{R}^3 except those with $y = 0$, $x \in [0, \infty)$ or with $x = 0$, $y = 0$, $z \in (-\infty, 0]$.
- (c) If v is a longitudinal unit tangent vector on the sphere $S^2(r)$ of radius r , then $d\phi(v) = 1$. If w points along a meridian through $p = (x, y, z) \in S^2(r)$,



then

$$d\theta(w_p) = \frac{1}{\sqrt{x^2 + y^2}}.$$

- (d) If θ and ϕ are taken to mean the restrictions of θ and ϕ to [certain portions of] S^2 , then

$$\sigma' = h d\theta \wedge d\phi,$$

where $h: S^2 \rightarrow \mathbb{R}$ is

$$h(x, y, z) = -\sqrt{x^2 + y^2} \quad (\text{the minus sign comes from the orientation}).$$

- (e) Conclude that

$$\sigma' = d(-\cos \phi d\theta).$$

(f) Let $r_2: \mathbb{R}^2 - \{0\} \rightarrow S^1$ be the retraction, so that $d\theta = r_2^* i^* \sigma$, for the form σ on \mathbb{R}^2 . Show that

$$r_2^* d\theta = d\theta.$$

If $\pi: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection, then the form $d\theta$ on [part of] \mathbb{R}^3 is just $\pi^* d\theta$, for the form $d\theta$ on [part of] \mathbb{R}^2 . Use this to show that

$$r^* d\theta = d\theta.$$

(g) Also prove this directly by using the result in part (c), and the fact that $r_*(v_p) = v_{r(p)}/|p|$ for v tangent to $S^2(|p|)$.

(h) Conclude that

$$\begin{aligned} d\Theta_3 &= r^* \sigma' = d(-\cos(\phi \circ r) d\theta) \\ &= d(-\cos \phi d\theta). \end{aligned}$$

(i) Similarly, express $d\Theta_n$ on $\mathbb{R}^n - \{0\}$ in terms of $d\Theta_{n-1}$ on $\mathbb{R}^{n-1} - \{0\}$.

20. Prove that a connected manifold is the union $U_1 \cup U_2 \cup U_3 \cup \dots$, where the U_i are coordinate neighborhoods, with $U_i \cap U_j \neq \emptyset$, and the sequence is eventually outside of any compact set.

21. Let $f: M^n \rightarrow N^n$ be a proper map between oriented n -manifolds such that $f_*: M_p \rightarrow N_{f(p)}$ is orientation preserving whenever p is a regular point. Show that if N is connected, then either f is onto N , or else all points are critical points of f .

22. (a) Show that a polynomial map $f: \mathbb{C} \rightarrow \mathbb{C}$, given by $f(z) = z^n + a_1 z^{n-1} + \dots + a_n$, is proper ($n \geq 1$).

(b) Let $f'(z) = nz^{n-1} + (n-1)a_1 z^{n-2} + \dots + a_{n-1}$. Show that we have $f'(z) = \lim_{w \rightarrow 0} [f(z+w) - f(z)]/w$, where w varies over complex numbers.

(c) Write $f(x+iy) = u(x, y) + iv(x, y)$ for real-valued functions u and v . Show that

$$\begin{aligned} f'(x+iy) &= \frac{\partial u}{\partial x}(x, y) + i \frac{\partial v}{\partial x}(x, y) \\ &= \frac{\partial v}{\partial y}(x, y) - i \frac{\partial u}{\partial y}(x, y). \end{aligned}$$

Hint: Choose w to be a real h , and then to be ih .

(d) Conclude that

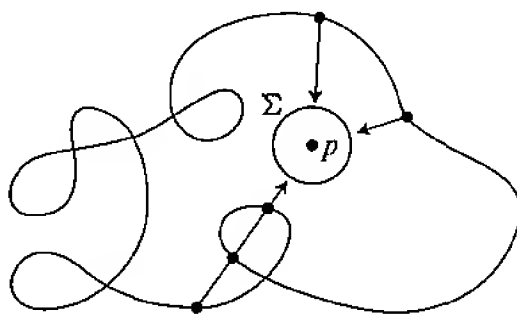
$$|f'(x+iy)|^2 = \det Df(x, y),$$

where f' is defined in part (b), while Df is the linear transformation defined for any differentiable $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$.

(e) Using Problem 21, give another proof of the Fundamental Theorem of Algebra.

(f) There is a still simpler argument, not using Problem 21 (which relies on many theorems of this chapter). Show directly that if $f: M \rightarrow N$ is proper, then the number of points in $f^{-1}(a)$ is a locally constant function on the set of regular values of f . Show that this set is connected for a polynomial $f: \mathbb{C} \rightarrow \mathbb{C}$, and conclude that f takes on all values.

23. Let $M^{n-1} \subset \mathbb{R}^n$ be a compact oriented manifold. For $p \in \mathbb{R}^n - M$, choose an $(n-1)$ -sphere Σ around p such that all points inside Σ are in $\mathbb{R}^n - M$. Let $r_p: \mathbb{R}^n - \{p\} \rightarrow \Sigma$ be the obvious retraction. Define the winding number $w(p)$ of M around p to be the degree of $r_p|_M$.

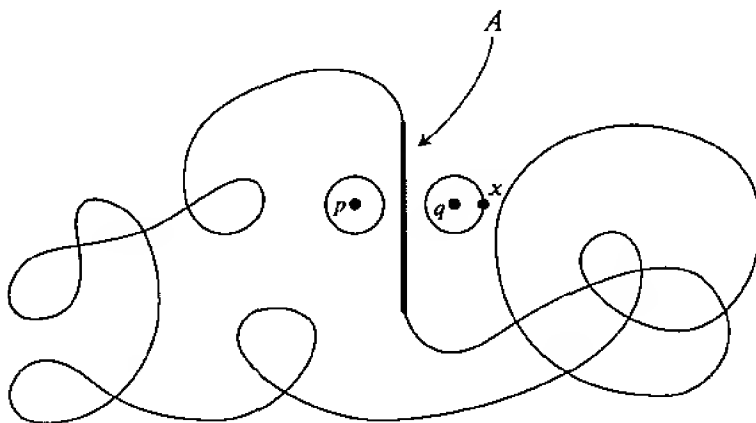


(a) Show that this definition agrees with that in Problem 3.

(b) Show that this definition does not depend on the choice of Σ .

(c) Show that w is constant in a neighborhood of p . Conclude that w is constant on each component of $\mathbb{R}^n - M$.

(d) Suppose M contains a portion A of an $(n-1)$ -plane. Let p and q be points



close to this plane, but on opposite sides. Show that $w(q) = w(p) \pm 1$. (Show that $r_q|M$ is homotopic to a map which equals $r_p|M$ on $M - A$ and which does not take any point of A onto the point x in the figure.)

(e) Show that, in general, if M is orientable, then $\mathbb{R}^n - M$ has at least 2 components. The next few Problems show how to prove the same result even if M is not orientable. More precise conclusions are drawn in Chapter 11.

24. Let M and N be compact n -manifolds, and let $f, g: M \rightarrow N$ be smoothly homotopic, by a smooth homotopy $H: M \times [0, 1] \rightarrow N$.

(a) Let $q \in N$ be a regular value of H . Let $\#f^{-1}(q)$ denote the (finite) number of points in $f^{-1}(q)$. Show that

$$\#f^{-1}(q) \equiv \#g^{-1}(q) \pmod{2}.$$

Hint: $H^{-1}(q)$ is a compact 1-manifold-with-boundary. The number of points in its boundary is clearly even. (This is one place where we use the stronger form of Sard's Theorem.)

(b) Show, more generally, that this result holds so long as q is a regular value of both f and g .

25. For two maps $f, g: M \rightarrow N$ we will write $f \simeq g$ to indicate that f is smoothly homotopic to g .

(a) If $f \simeq g$, then there is a smooth homotopy $H': M \times [0, 1] \rightarrow N$ such that

$$H'(p, t) = f(p) \quad \text{for } t \text{ in a neighborhood of } 0,$$

$$H'(p, t) = g(p) \quad \text{for } t \text{ in a neighborhood of } 1.$$

(b) \simeq is an equivalence relation.

26. If f is smoothly homotopic to g by a smooth homotopy H such that $p \mapsto H(p, t)$ is a diffeomorphism for each t , we say that f is **smoothly isotopic** to g .

(a) Being smoothly isotopic is an equivalence relation.

(b) Let $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ be a C^∞ function which is positive on the interior of the unit ball, and 0 elsewhere. For $p \in S^{n-1}$, let $H: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfy

$$\frac{\partial H(t, x)}{\partial t} = \phi(H(t, x)) \cdot p$$

$$H(0, x) = x.$$

(Each solution is defined for all t , by Theorem 5-6.) Show that each $x \mapsto H(t, x)$ is a diffeomorphism, which is smoothly isotopic to the identity, and leaves all points outside the unit ball fixed.

- (c) Show that by choosing suitable p and t we can make $H(t, 0)$ be any point in the interior of the unit ball.
- (d) If M is connected and $p, q \in M$, then there is a diffeomorphism $f: M \rightarrow M$ such that $f(p) = q$ and f is smoothly isotopic to the identity.
- (e) Use part (d) to give an alternate proof of *Step 3* of Theorem 9.
- (f) If M and N are compact n -manifolds, and $f: M \rightarrow N$, then for regular values $q_1, q_2 \in N$ we have

$$\#f^{-1}(q_1) \equiv \#f^{-1}(q_2) \pmod{2}$$

(where $\#f^{-1}(q)$ is defined in Problem 24). This number is called the mod 2 degree of f .

- (g) By replacing “degree” with “mod 2 degree” in Problem 23, show that if $M \subset \mathbb{R}^n$ is a compact $(n-1)$ -manifold, then $\mathbb{R}^n - M$ has at least 2 components.

27. Let $\{X^t\}$ be a C^∞ family of C^∞ vector fields on a compact manifold M . (To be more precise, suppose X is a C^∞ vector field on $M \times [0, 1]$; then $X^t(p)$ will denote $\pi_{M*} X_{(p,t)}$.) From the addendum to Chapter 5, and the argument which was used in the proof of Theorem 5-6, it follows that there is a C^∞ family $\{\phi_t\}$ of diffeomorphisms of M [not necessarily a 1-parameter group], with $\phi_0 = \text{identity}$, which is generated by $\{X^t\}$, i.e., for any C^∞ function $f: M \rightarrow \mathbb{R}$ we have

$$(X^t f)(p) = \lim_{h \rightarrow 0} \frac{f(\phi_{t+h}(p)) - f(\phi_t(p))}{h}.$$

For a family ω_t of k -forms on M we define the k -form

$$\dot{\omega}_t = \lim_{h \rightarrow 0} \frac{\omega_{t+h} - \omega_t}{h}.$$

- (a) Show that for $\eta(t) = \phi_t^* \omega_t$ we have

$$\dot{\eta}_t = \phi_t^*(L_{X^t} \omega_t + \dot{\omega}_t).$$

- (b) Let ω_0 and ω_1 be nowhere zero n -forms on a compact oriented n -manifold M , and define

$$\omega_t = (1-t)\omega_0 + t\omega_1.$$

Show that the family ϕ_t of diffeomorphisms generated by $\{X^t\}$ satisfies

$$\phi_t^* \omega_t = \omega_0 \quad \text{for all } t$$

if and only if

$$L_{X^t} \omega_t = \omega_0 - \omega_1.$$

(c) Using Problem 7-18, show that this holds if and only if

$$d(X^t \lrcorner \omega_t) = \omega_0 - \omega_1.$$

(d) Suppose that $\int_M \omega_0 = \int_M \omega_1$, so that $\omega_0 - \omega_1 = d\lambda$ for some λ . Show that there is a diffeomorphism $f_1: M \rightarrow M$ such that $\omega_0 = f_1^* \omega_1$.

28. Let $f: M^k \rightarrow \mathbb{R}^n$ and $g: N^l \rightarrow \mathbb{R}^n$ be C^∞ maps, where M and N are compact oriented manifolds, $n = k + l + 1$, and $f(M) \cap g(N) = \emptyset$. Define

$$\alpha_{f,g}: M \times N \rightarrow S^{n-1} \subset \mathbb{R}^n - \{0\}$$

by

$$\alpha_{f,g}(p, q) = r(g(q) - f(p)) = \frac{g(q) - f(p)}{|g(q) - f(p)|}.$$

We define the linking number of f and g to be

$$\ell(f, g) = \deg \alpha_{f,g},$$

where $M \times N$ is oriented as in Problem 18.

(a) $\ell(f, g) = (-1)^{kl+1} \ell(g, f)$.

(b) Let $H: M \times [0, 1] \rightarrow \mathbb{R}^n$ and $K: N \times [0, 1] \rightarrow \mathbb{R}^n$ be smooth homotopies with

$$\begin{aligned} H(p, 0) &= f(p) & K(q, 0) &= g(q) \\ H(p, 1) &= \tilde{f}(p) & K(q, 1) &= \tilde{g}(q) \end{aligned}$$

such that

$$\{H(p, t) : p \in M\} \cap \{K(q, t) : q \in N\} = \emptyset \quad \text{for every } t.$$

Show that

$$\ell(f, g) = \ell(\tilde{f}, \tilde{g}).$$

(c) For $f, g: S^1 \rightarrow \mathbb{R}^3$ show that

$$\ell(f, g) = \frac{-1}{4\pi} \int_0^1 \int_0^1 \frac{A(u, v)}{[r(u, v)]^3} du dv,$$

where

$$r(u, v) = |g(v) - f(u)|$$

$$A(u, v) = \det \begin{pmatrix} (f^1)'(u) & (f^2)'(u) & (f^3)'(u) \\ (g^1)'(v) & (g^2)'(v) & (g^3)'(v) \\ g^1(v) - f^1(u) & g^2(v) - f^2(u) & g^3(v) - f^3(u) \end{pmatrix}$$

(the factor $1/4\pi$ comes from the fact that $\int_{S^2} \sigma' = 4\pi$ [Problem 9-14]).

(d) Show that $\ell(f, g) = 0$ if f and g both lie in the same plane (first do it for (x, y) -plane). The next problem shows how to determine $\ell(f, g)$ without calculating.

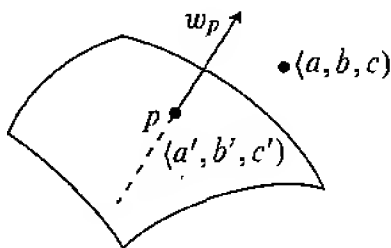
29. (a) For $(a, b, c) \in \mathbb{R}^3$ define

$$d\Theta_{(a,b,c)} = \frac{(x-a)dy \wedge dz - (y-b)dx \wedge dz + (z-c)dx \wedge dy}{[(x-a)^2 + (y-b)^2 + (z-c)^2]^{3/2}}.$$

For a compact oriented 2-manifold-with-boundary $M \subset \mathbb{R}^3$ and $(a, b, c) \notin M$, let

$$\Omega(a, b, c) = \int_M d\Theta_{(a,b,c)}.$$

Let (a, b, c) and (a', b', c') be points close to $p \in M$, on opposite sides of M . Suppose (a, b, c) is on the same side as a vector $w_p \in \mathbb{R}^3_p - M_p$ for which the



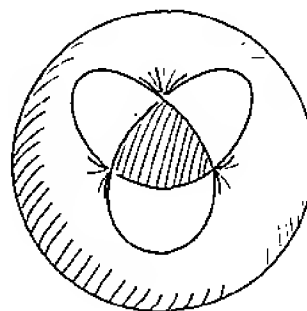
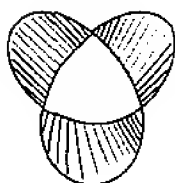
triple $w_p, (v_1)_p, (v_2)_p$ is positively oriented in \mathbb{R}^3_p when $(v_1)_p, (v_2)_p$ is positively oriented in M_p . Show that

$$\lim_{\substack{(a,b,c) \rightarrow p \\ (a',b',c') \rightarrow p}} \Omega(a, b, c) - \Omega(a', b', c') = -4\pi.$$

Hint: First show that if $M = \partial N$, then $\Omega(a, b, c) = -4\pi$ for $(a, b, c) \in N - M$ and $\Omega(a, b, c) = 0$ for $(a, b, c) \notin N$.

(b) Let $f: S^1 \rightarrow \mathbb{R}^3$ be an imbedding such that $f(S^1) = \partial M$ for some compact oriented 2-manifold-with-boundary M . (An M with this property always exists. See Fort, *Topology of 3-Manifolds*, pg. 138.) Let $g: S^1 \rightarrow \mathbb{R}^3$ and suppose

The figure on the left shows a *non-orientable* surface whose boundary is the “trefoil” knot,



but the surface on the right—including the hemisphere behind the plane of the paper—is orientable.

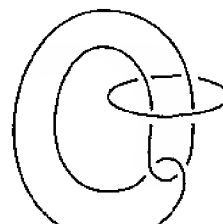
that when $g(t) = p \in M$ we have $dg/dt \notin M_p$. Let n^+ be the number of intersections where dg/dt points in the same direction as the vector w_p of part (a), and n^- the number of other intersections. Show that

$$n = n^+ - n^- = \frac{-1}{4\pi} \int_{S^1} g^*(d\Omega).$$

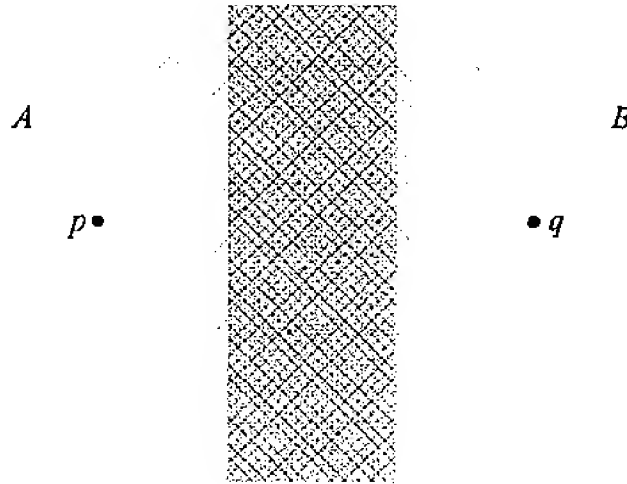
(c) Show that

$$\begin{aligned} \frac{\partial \Omega}{\partial a}(a, b, c) &= \int_{S^1} f^* \left(\frac{(y-b) dz - (z-c) dy}{|(x, y, z)|^3} \right) \\ \frac{\partial \Omega}{\partial b}(a, b, c) &= \int_{S^1} f^* \left(\frac{(z-c) dx - (x-a) dz}{|(x, y, z)|^3} \right) \\ \frac{\partial \Omega}{\partial c}(a, b, c) &= \int_{S^1} f^* \left(\frac{(x-a) dy - (y-b) dx}{|(x, y, z)|^3} \right). \end{aligned}$$

(d) Show that $n = \ell(f, g)$. Compute $\ell(f, g)$ for the pairs shown below.



30. (a) Let $p, q \in \mathbb{R}^n$ be distinct. Choose open sets $A, B \subset \mathbb{R}^n - \{p, q\}$ so that A and B are diffeomorphic to $\mathbb{R}^n - \{0\}$, and $A \cap B$ is diffeomorphic to \mathbb{R}^n . Using an argument similar to that in the proof of Theorem 16, show that



$H^k(\mathbb{R}^n - \{p, q\}) = 0$ for $0 < k < n - 1$, and that $H^{n-1}(\mathbb{R}^n - \{p, q\})$ has dimension 2.

(b) Find the de Rham cohomology vector spaces of $\mathbb{R}^n - F$ where $F \subset \mathbb{R}^n$ is a finite set.

31. We define the cup product $\cup: H^k(M) \times H^l(M) \rightarrow H^{k+l}(M)$ by

$$[\omega] \cup [\eta] = [\omega \wedge \eta].$$

(a) Show that \cup is well-defined, i.e., $\omega \wedge \eta$ is exact if ω is exact and η is closed.

(b) Show that \cup is bilinear.

(c) If $\alpha \in H^k(M)$ and $\beta \in H^l(M)$, then $\alpha \cup \beta = (-1)^{kl} \beta \cup \alpha$.

(d) If $f: M \rightarrow N$, and $\alpha \in H^k(N)$, $\beta \in H^l(N)$, then

$$f^*(\alpha \cup \beta) = f^*\alpha \cup f^*\beta.$$

(e) The cross-product $\times: H^k(M) \times H^l(N) \rightarrow H^{k+l}(M \times N)$ is defined by

$$[\omega] \times [\eta] = [\pi_M^*\omega \wedge \pi_N^*\eta].$$

Show that \times is well-defined, and that

$$\alpha \times \beta = \pi_M^*\alpha \cup \pi_N^*\beta.$$

(f) If $\Delta: M \rightarrow M \times M$ is the “diagonal map”, given by $\Delta(p) = (p, p)$, show that

$$\alpha \cup \beta = \Delta^*(\alpha \times \beta).$$

32. On the n -dimensional torus

$$T^n = \underbrace{S^1 \times \cdots \times S^1}_{n \text{ times}}$$

let $d\theta^i$ denote $\pi_i^*d\theta$, where $\pi_i: T^n \rightarrow S^1$ is projection on the i^{th} factor.

(a) Show that all $d\theta^{i_1} \wedge \cdots \wedge d\theta^{i_k}$ represent different elements of $H^k(T^n)$, by finding submanifolds of T^n over which they have different integrals. Hence $\dim H^k(T^n) \geq \binom{n}{k}$. Equality is proved in the Problems for Chapter 11.

(b) Show that every map $f: S^n \rightarrow T^n$ has degree 0. *Hint:* Use Problem 25.

CHAPTER 9

RIEMANNIAN METRICS

In previous chapters we have exploited nearly every construction associated with vector spaces, and thus with bundles, but there has been one notable exception—we have never mentioned inner products. The time has now come to make use of this neglected tool.

An **inner product** on a vector space V over a field F is a bilinear function from $V \times V$ to F , denoted by $(v, w) \mapsto (v, w)$, which is **symmetric**,

$$(v, w) = (w, v),$$

and **non-degenerate**: if $v \neq 0$, then there is some $w \neq 0$ such that

$$(w, v) \neq 0.$$

For us, the field F will always be \mathbb{R} .

For each r with $0 \leq r \leq n$, we can define an inner product $(\ , \)_r$ on \mathbb{R}^n by

$$(a, b)_r = \sum_{i=1}^r a^i b^i - \sum_{i=r+1}^n a^i b^i;$$

this is non-degenerate because if $a \neq 0$, then

$$\langle (a^1, \dots, a^n), (a^1, \dots, a^r, -a^{r+1}, \dots, -a^n) \rangle_r = \sum_{i=1}^n (a^i)^2 > 0.$$

In particular, for $r = n$ we obtain the “usual inner product”, $\langle \ , \ \rangle$ on \mathbb{R}^n ,

$$\langle a, b \rangle = \sum_{i=1}^n a^i b^i.$$

For this inner product we have $\langle a, a \rangle > 0$ for any $a \neq 0$. In general, a symmetric bilinear function $(\ , \)$ is called **positive definite** if

$$(v, v) > 0 \quad \text{for all } v \neq 0.$$

A positive definite bilinear function $(\ , \)$ is clearly non-degenerate, and consequently an inner product.

Notice that an inner product $(\ , \)$ on V is an element of $\mathcal{T}^2(V)$, so if $f: W \rightarrow V$ is a linear transformation, then $f^*(\ , \)$ is a symmetric bilinear function on W . This symmetric bilinear function may be degenerate even if f is one-one, e.g., if $(\ , \)$ is defined on \mathbb{R}^2 by

$$(a, b) = a^1 b^1 - a^2 b^2,$$

and $f: \mathbb{R} \rightarrow \mathbb{R}^2$ is

$$f(a) = (a, a).$$

However, $f^*(\ , \)$ is clearly non-degenerate if f is an isomorphism *onto* V . Also, if $(\ , \)$ is positive definite, then $f^*(\ , \)$ is positive definite if and only if f is one-one.

For any basis v_1, \dots, v_n of V , with corresponding dual basis v^*_1, \dots, v^*_n , we can write

$$(\ , \) = \sum_{i,j=1}^n g_{ij} v^*_i \otimes v^*_j.$$

In this expression,

$$g_{ij} = (v_i, v_j),$$

so symmetry of $(\ , \)$ implies that the matrix $\langle g_{ij} \rangle$ is symmetric,

$$g_{ij} = g_{ji}.$$

The matrix $\langle g_{ij} \rangle$ has another important interpretation. Since an inner product $(\ , \)$ is linear in the second argument, we can define a linear functional $\phi_v \in V^*$, for each $v \in V$, by

$$\phi_v(w) = (v, w).$$

Since $(\ , \)$ is linear in the first argument, the map $v \mapsto \phi_v$ is a linear transformation from V to V^* . Non-degeneracy of $(\ , \)$ implies that $\phi_v \neq 0$ if $v \neq 0$. Thus, if V is finite dimensional, an inner product $(\ , \)$ gives us an isomorphism $\alpha: V \rightarrow V^*$, with

$$(v, w) = \alpha(v)(w).$$

Clearly, the matrix $\langle g_{ij} \rangle$ is just the matrix of $\alpha: V \rightarrow V^*$ with respect to the bases $\{v_i\}$ for V and $\{v^*_i\}$ for V^* . Thus, non-degeneracy of $(\ , \)$ is equivalent to the condition that

$$\langle g_{ij} \rangle \text{ is non-singular, } \det(g_{ij}) \neq 0.$$

Positive definiteness of $(\ , \)$ corresponds to the more complicated condition that the matrix $\langle g_{ij} \rangle$ be “positive definite”, meaning that

$$\sum_{i=1}^n g_{ij} a^i a^j > 0 \quad \text{for all } a_1, \dots, a_n \text{ with at least one } a^i \neq 0.$$

Given any *positive definite* inner product $\langle \cdot, \cdot \rangle$ on V we define the associated norm $\| \cdot \|$ by

$$\|v\| = \sqrt{\langle v, v \rangle} \quad (\text{the positive square root is to be taken}).$$

In \mathbb{R}^n we denote the norm corresponding to $\langle \cdot, \cdot \rangle$ simply by

$$|a| = \sqrt{\langle a, a \rangle} = \sqrt{\sum_{i=1}^n (a^i)^2}.$$

The principal properties of $\| \cdot \|$ are the following

1. THEOREM. For all $v, w \in V$ we have

- (1) $\|av\| = |a| \cdot \|v\|$.
- (2) $|\langle v, w \rangle| \leq \|v\| \cdot \|w\|$, with equality if and only if v and w are linearly dependent (Schwarz inequality).
- (3) $\|v + w\| \leq \|v\| + \|w\|$ (Triangle inequality).

PROOF. (1) is trivial.

(2) If v and w are linearly dependent, equality clearly holds. If not, then $0 \neq \lambda v - w$ for all $\lambda \in \mathbb{R}$, so

$$\begin{aligned} 0 < \|\lambda v - w\|^2 &= \langle \lambda v - w, \lambda v - w \rangle \\ &= \lambda^2 \|v\|^2 - 2\lambda \langle v, w \rangle + \|w\|^2. \end{aligned}$$

So the right side is a quadratic equation in λ with no real solution, and its discriminant must be negative. Thus

$$4\langle v, w \rangle^2 - 4\|v\|^2\|w\|^2 < 0.$$

$$\begin{aligned} (3) \quad \|v + w\|^2 &= \langle v + w, v + w \rangle \\ &= \|v\|^2 + \|w\|^2 + 2\langle v, w \rangle \\ &\leq \|v\|^2 + \|w\|^2 + 2\|v\| \cdot \|w\| \quad \text{by (2)} \\ &= (\|v\| + \|w\|)^2. \quad \spadesuit \end{aligned}$$

The function $\| \cdot \|$ has certain unpleasant properties—for example, the function $| \cdot |$ on \mathbb{R}^n is not differentiable at $0 \in \mathbb{R}^n$ —which do not arise for the function $\| \cdot \|^2$. This latter function is a “quadratic function” on V —in terms of a basis $\{v_i\}$ for V it can be written as a “homogeneous polynomial of degree 2” in the components,

$$\left\| \sum_{i=1}^n a^i v^i \right\|^2 = \sum_{i,j=1}^n g_{ij} a^i a^j.$$

More succinctly,

$$\| \|^2 = \sum_{i,j=1}^n g_{ij} v_i^* \cdot v_j^*.$$

An invariant definition of a quadratic function can be obtained (Problem 1) from the following observation.

2. THEOREM (POLARIZATION IDENTITY). If $\| \cdot \|$ is the norm associated to an inner product $\langle \cdot, \cdot \rangle$ on V , then

$$(1) \quad \langle v, w \rangle = \frac{1}{2} [\|v + w\|^2 - \|v\|^2 - \|w\|^2]$$

$$(2) \quad \langle v, w \rangle = \frac{1}{4} [\|v + w\|^2 - \|v - w\|^2].$$

PROOF. Compute. ♦

Theorem 2 shows that two inner products which induce the same norm are themselves equal. Similarly, if $f: V \rightarrow V$ is norm preserving, that is, $\|f(v)\| = \|v\|$ for all $v \in V$, then f is also inner product preserving, that is, $\langle f(v), f(w) \rangle = \langle v, w \rangle$ for all $v, w \in V$.

We will now see that, “up to isomorphism”, there is only one positive definite inner product.

3. THEOREM. If $\langle \cdot, \cdot \rangle$ is a positive definite inner product on an n -dimensional vector space V , then there is a basis v_1, \dots, v_n for V such that $\langle v_i, v_j \rangle = \delta_{ij}$. (Such a basis is called **orthonormal** with respect to $\langle \cdot, \cdot \rangle$.) Consequently, there is an isomorphism $f: \mathbb{R}^n \rightarrow V$ such that

$$\langle a, b \rangle = \langle f(a), f(b) \rangle, \quad a, b \in \mathbb{R}^n.$$

In other words,

$$f^* \langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle.$$

PROOF. Let w_1, \dots, w_n be any basis for V . We obtain the desired basis by applying the “Gram-Schmidt orthonormalization process” to this basis:

Since $w_1 \neq 0$, we can define

$$v_1 = \frac{w_1}{\|w_1\|},$$

and clearly $\|v_1\| = 1$. Suppose that we have constructed v_1, \dots, v_k so that

$$\langle v_i, v_j \rangle = \delta_{ij} \quad 1 \leq i, j \leq k$$

and

$$\text{span } v_1, \dots, v_k = \text{span } w_1, \dots, w_k.$$

Then w_{k+1} is linearly independent of v_1, \dots, v_k . Let

$$w'_{k+1} = w_{k+1} - \langle v_1, w_{k+1} \rangle v_1 - \dots - \langle v_k, w_{k+1} \rangle v_k \neq 0.$$

It is easy to see that

$$\langle w'_{k+1}, v_i \rangle = 0 \quad i = 1, \dots, k.$$

So we can define

$$v_{k+1} = \frac{w'_{k+1}}{\|w'_{k+1}\|},$$

and continue inductively. ♦

A positive definite inner product $\langle \cdot, \cdot \rangle$ on V is sometimes called a *Euclidean metric* on V . This is because we obtain a metric ρ on V by defining

$$\rho(v, w) = \|v - w\|.$$

The “triangle inequality” (Theorem 1(3)) shows that this is indeed a metric. We also call $\|v\|$ the **length** of v .

We have only one more algebraic trick to play. Recall that an inner product $\langle \cdot, \cdot \rangle$ on V provides an isomorphism $\alpha: V \rightarrow V^*$ with

$$\alpha(v)(w) = \langle v, w \rangle.$$

Using the natural isomorphism $i: V \rightarrow V^{**}$, defined by

$$i(v)(\lambda) = \lambda(v),$$

we obtain an isomorphism

$$\beta: V^* \xrightarrow{\alpha^{-1}} V \xrightarrow{i} (V^*)^*.$$

We can now use β to define a bilinear function $\langle \cdot, \cdot \rangle^*$ on V^* by

$$\langle \lambda, \mu \rangle^* = \beta(\lambda)(\mu) = i\alpha^{-1}(\lambda)(\mu) = \mu(\alpha^{-1}(\lambda)).$$

Now, the symmetry of $\langle \cdot, \cdot \rangle$ can be expressed by the equation

$$\alpha(v)(w) = \alpha(w)(v).$$

Letting

$$\alpha(v) = \lambda, \quad \alpha(w) = \mu,$$

this can be written

$$\lambda(\alpha^{-1}(\mu)) = \mu(\alpha^{-1}(\lambda)),$$

which shows that $\langle \cdot, \cdot \rangle^*$ is also symmetric,

$$(\mu, \lambda)^* = (\lambda, \mu)^*.$$

Consequently $\langle \cdot, \cdot \rangle^*$ is an inner product on the dual space V^* (in fact, the one which produces β).

To see what this all means, choose a basis $\{v_i\}$ for V , let $\{v_i^*\}$ be the dual basis for V^* , and let

$$\langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} v_i^* \otimes v_j^*.$$

Then

(g_{ij}) is the matrix of $\alpha: V \rightarrow V^*$ with respect to $\{v_i\}$ and $\{v_i^*\}$

so

$(g_{ij})^{-1}$ is the matrix of $\alpha^{-1}: V^* \rightarrow V$ with respect to $\{v_i^*\}$ and $\{v_i\}$

so

$(g_{ij})^{-1}$ is the matrix of $\beta: V^* \rightarrow V^{**}$ with respect to $\{v_i^*\}$ and $\{v_i^{**}\}$.

Thus, if we let g^{ij} be the entries of the inverse matrix, $(g^{ij}) = (g_{ij})^{-1}$, so that

$$\sum_{k=1}^n g^{ik} g_{kj} = \delta_j^i,$$

then

$$\begin{aligned} \langle \cdot, \cdot \rangle^* &= \sum_{i,j=1}^n g^{ij} v_i^{**} \otimes v_j^{**} \\ &= \sum_{i,j=1}^n g^{ij} v_i \otimes v_j, \quad \text{if we consider } v_i \in V^{**}. \end{aligned}$$

One can check directly (Problem 9), without the invariant definition, that this equation defines $\langle \cdot, \cdot \rangle^*$ independently of the choice of basis.

Notice that if (\cdot, \cdot) is positive definite, so that

$$\alpha(v)(v) > 0 \quad \text{for } v \neq 0,$$

then, letting $\alpha(v) = \lambda$, we have

$$\lambda(\alpha^{-1}(\lambda)) = \beta(\lambda)(\lambda) > 0 \quad \text{for } \lambda \neq 0,$$

so $(\cdot, \cdot)^*$ is also positive definite. This can also be checked directly from the definition in terms of a basis. In the positive definite case, the simplest way to describe $(\cdot, \cdot)^*$ is as follows: The basis v_1^*, \dots, v_n^* of V^* is orthonormal with respect to $(\cdot, \cdot)^*$ if and only if v_1, \dots, v_n is orthonormal with respect to (\cdot, \cdot) .

Similar tricks can be used (Problem 4) to produce an inner product on all the vector spaces $\mathcal{T}^k(V)$, $\mathcal{T}_k(V) = \mathcal{T}^k(V^*)$, and $\Omega^k(V)$. However, we are interested in only one case, which we will not describe in a completely invariant way. The vector space $\Omega^n(V)$ is 1-dimensional, so to produce an inner product on it, we need only describe which two elements, ω and $-\omega$, will have length 1. Let v_1, \dots, v_n and w_1, \dots, w_n be two bases of V which are orthonormal with respect to (\cdot, \cdot) . If we write

$$w_i = \sum_{j=1}^n \alpha_{ji} v_j,$$

then

$$\begin{aligned} \delta_{ij} = \langle w_i, w_j \rangle &= \left\langle \sum_{k=1}^n \alpha_{ki} v_k, \sum_{l=1}^n \alpha_{lj} v_l \right\rangle = \sum_{k,l=1}^n \alpha_{ki} \alpha_{lj} \langle v_k, v_l \rangle \\ &= \sum_{k=1}^n \alpha_{ki} \alpha_{kj}. \end{aligned}$$

So the transpose matrix A^t of $A = (\alpha_{ij})$ satisfies $A \cdot A^t = I$, which implies that $\det A = \pm 1$. It follows from Theorem 7-5 that for any $\omega \in \Omega^n(V)$ we have

$$\omega(v_1, \dots, v_n) = \pm \omega(w_1, \dots, w_n).$$

It clearly follows that

$$v_1^* \wedge \dots \wedge v_n^* = \pm w_1^* \wedge \dots \wedge w_n^*.$$

We have thus distinguished two elements of $\Omega^n(V)$; they are both of the form $v_1^* \wedge \dots \wedge v_n^*$ for $\{v_i\}$ an orthonormal basis of V . We will call these two elements

the elements of norm 1 in $\Omega^n(V)$. If we also have an orientation μ , then we can further distinguish the one which is positive when applied to any (v_1, \dots, v_n) with $[v_1, \dots, v_n] = \mu$; we will call it the **positive element of norm 1** in $\Omega^n(V)$.

To express the elements of norm 1 in terms of an arbitrary basis w_1, \dots, w_n , we choose an orthonormal basis v_1, \dots, v_n and write

$$w_i = \sum_{j=1}^n \alpha_{ji} v_j.$$

Problem 7-9 implies that

$$\det(\alpha_{ij}) w^*_1 \wedge \cdots \wedge w^*_n = v^*_1 \wedge \cdots \wedge v^*_n.$$

If we write

$$\langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} w^*_i \otimes w^*_j,$$

then

$$\begin{aligned} g_{ij} = \langle w_i, w_j \rangle &= \left\langle \sum_{k=1}^n \alpha_{ki} v_k, \sum_{l=1}^n \alpha_{lj} v_l \right\rangle \\ &= \sum_{k=1}^n \alpha_{ki} \alpha_{kj}, \end{aligned}$$

so if $A = (\alpha_{ij})$, then

$$\det(g_{ij}) = \det(A^t \cdot A) = (\det A)^2.$$

In particular, $\det(g_{ij})$ is always positive. Consequently, the elements of norm 1 in $\Omega^n(V)$ are

$$\pm \sqrt{\det(g_{ij})} w^*_1 \wedge \cdots \wedge w^*_n \quad g_{ij} = (w_i, w_j).$$

We now apply our new tool to vector bundles. If $\xi = \pi: E \rightarrow B$ is a vector bundle, we define a **Riemannian metric** on ξ to be a function $\langle \cdot, \cdot \rangle$ which assigns to each $p \in B$ a positive definite inner product $\langle \cdot, \cdot \rangle_p$ on $\pi^{-1}(p)$, and which is continuous in the sense that for any two continuous sections $s_1, s_2: B \rightarrow E$, the function

$$\langle s_1, s_2 \rangle = p \mapsto \langle s_1(p), s_2(p) \rangle_p$$

is also continuous. If ξ is a C^∞ vector bundle over a C^∞ manifold we can also speak of C^∞ Riemannian metrics.

[Another approach to the definition can be given. Let $Euc(V)$ be the set of all positive definite inner products on V . If we replace each $\pi^{-1}(p)$ by $Euc(\pi^{-1}(p))$, and let

$$Euc(\xi) = \bigcup_{p \in B} Euc(\pi^{-1}(p)),$$

then a Riemannian metric on ξ can be defined to be a section of $Euc(\xi)$. The only problem is that $Euc(V)$ is not a vector space; the new object $Euc(\xi)$ that we obtain is not a vector bundle at all, but an instance of a more general structure, a fibre bundle.]

4. THEOREM. Let $\xi = \pi: E \rightarrow M$ be a $[C^\infty]$ k -plane bundle over a C^∞ manifold M . Then there is a $[C^\infty]$ Riemannian metric on ξ .

PROOF. There is an open locally finite cover \mathcal{O} of M by sets U for which there exists $[C^\infty]$ trivializations

$$t_U: \pi^{-1}(U) \rightarrow U \times \mathbb{R}^k.$$

On $U \times \mathbb{R}^k$, there is an obvious Riemannian metric,

$$\langle (p, a), (p, b) \rangle_p = \langle a, b \rangle.$$

For $v, w \in \pi^{-1}(p)$, define

$$\langle v, w \rangle_p^U = \langle t_U(v), t_U(w) \rangle_p.$$

Then $\langle \cdot, \cdot \rangle^U$ is a $[C^\infty]$ Riemannian metric for $\xi|U$. Let $\{\phi_U\}$ be a partition of unity subordinate to \mathcal{O} . We define $\langle \cdot, \cdot \rangle$ by

$$\langle v, w \rangle_p = \sum_{U \in \mathcal{O}} \phi_U(p) \langle v, w \rangle_p^U \quad v, w \in \pi^{-1}(p).$$

Then $\langle \cdot, \cdot \rangle$ is continuous $[C^\infty]$ and each $\langle \cdot, \cdot \rangle_p$ is a symmetric bilinear function on $\pi^{-1}(p)$. To show that it is positive definite, note that

$$\langle v, v \rangle_p = \sum_{U \in \mathcal{O}} \phi_U(p) \langle v, v \rangle_p^U;$$

each $\phi_U(p) \langle v, v \rangle_p^U \geq 0$, and for some U strict inequality holds. ♦

[The same argument shows that any vector bundle over a paracompact space has a Riemannian metric.]

Notice that the argument in the final step would not work if we had merely picked non-degenerate inner products $\langle \cdot, \cdot \rangle^U$. In fact (Problem 7), there is no $\langle \cdot, \cdot \rangle$ on TS^2 which gives a symmetric bilinear function on each S^2_p which is not positive definite or negative definite but is still non-degenerate.

As an application of Theorem 4, we settle some questions which have till now remained unanswered.

5. COROLLARY. If $\xi = \pi : E \rightarrow M$ is a k -plane bundle, then $\xi \simeq \xi^*$.

PROOF. Let $\langle \cdot, \cdot \rangle$ be a Riemannian metric for ξ . Then for each $p \in M$, we have an isomorphism

$$\alpha_p : \pi^{-1}(p) \rightarrow [\pi^{-1}(p)]^*$$

defined by

$$\alpha_p(v)(w) = \langle v, w \rangle_p \quad v, w \in \pi^{-1}(p).$$

Continuity of $\langle \cdot, \cdot \rangle$ implies that the union of all α_p is a homeomorphism from E to $E' = \bigcup_{p \in M} [\pi^{-1}(p)]^*$. ♦

6. COROLLARY. If $\xi = \pi : E \rightarrow M$ is a 1-plane bundle, then ξ is trivial if and only if ξ is orientable.

PROOF. The “only if” part is trivial. If ξ has an orientation μ and $\langle \cdot, \cdot \rangle$ is a Riemannian metric on M then there is a unique

$$s(p) \in \pi^{-1}(p)$$

with

$$\langle s(p), s(p) \rangle_p = 1, \quad [s(p)] = \mu_p.$$

Clearly s is a section; we then define an equivalence $f : E \rightarrow M \times \mathbb{R}$ by

$$f(\lambda s(p)) = (p, \lambda).$$

ALTERNATIVE PROOF. We know (see the discussion after Theorem 7-9) that if ξ is orientable, then there is a nowhere 0 section of

$$\Omega^1(\xi) = \xi^*,$$

so that ξ^* is trivial. But $\xi \simeq \xi^*$. ♦

All these considerations take on special significance when our bundle is the tangent bundle TM of a C^∞ manifold M . In this case, a C^∞ Riemannian metric $\langle \cdot, \cdot \rangle$ for TM , which gives a positive definite inner product $\langle \cdot, \cdot \rangle_p$ on

each M_p , is called a **Riemannian metric on M** . If (x, U) is a coordinate system on M , then on U we can write our Riemannian metric $\langle \cdot, \cdot \rangle$ as

$$\langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j,$$

where the C^∞ functions g_{ij} satisfy $g_{ij} = g_{ji}$, since $\langle \cdot, \cdot \rangle$ is symmetric, and $\det(g_{ij}) > 0$ since $\langle \cdot, \cdot \rangle$ is positive definite. A Riemannian metric $\langle \cdot, \cdot \rangle$ on M is, of course, a covariant tensor of order 2. So for every C^∞ map $f: N \rightarrow M$ there is a covariant tensor $f^*\langle \cdot, \cdot \rangle$ on N , which is clearly symmetric; it is a Riemannian metric on N if and only if f is an immersion (f_{*p} is one-one for all $p \in N$).

The Riemannian metric $\langle \cdot, \cdot \rangle^*$, which $\langle \cdot, \cdot \rangle$ induces on the dual bundle T^*M , is a contravariant tensor of order 2, and we can write it as

$$\langle \cdot, \cdot \rangle^* = \sum_{i,j=1}^n g^{ij} \frac{\partial}{\partial x^i} \otimes \frac{\partial}{\partial x^j}.$$

Our discussion of inner products induced on V^* shows that for each p , the matrix $(g^{ij}(p))$ is the inverse of the matrix $(g_{ij}(p))$; thus

$$\sum_{k=1}^n g_{ik} g^{kj} = \delta_i^j.$$

Similarly, for each $p \in M$ the Riemannian metric $\langle \cdot, \cdot \rangle$ on M determines two elements of $\Omega^n(M_p)$, the elements of norm 1. We have seen that they can be written

$$\pm \sqrt{\det(g_{ij}(p))} dx^1(p) \wedge \cdots \wedge dx^n(p).$$

If M has an orientation μ , then μ_p allows us to pick out the positive element of norm 1, and we obtain an n -form on M ; if $x: U \rightarrow \mathbb{R}^n$ is *orientation preserving*, then on U this form can be written

$$\sqrt{\det(g_{ij})} dx^1 \wedge \cdots \wedge dx^n.$$

Even if M is not orientable, we obtain a “volume element” on M , as defined in Chapter 8; in a coordinate system (x, U) it can be written as

$$\sqrt{\det(g_{ij})} |dx^1 \wedge \cdots \wedge dx^n|.$$

This volume element is denoted by dV , even though it is usually not d of anything (even when M is orientable and it can be considered to be an n -form),

and is called the volume element determined by (\cdot, \cdot) . We can then define the volume of M as

$$\int_M dV.$$

This certainly makes sense if M is compact, and in the non-compact case (see Problem 8-10) it either converges to a definite number, or becomes arbitrarily large over compact subsets of M , in which case we say that M has “infinite volume”.

If M is an n -dimensional manifold (-with-boundary) in \mathbb{R}^n , with the “usual Riemannian metric”

$$(\cdot, \cdot) = \sum_{i=1}^n dx^i \otimes dx^i,$$

then $g_{ij} = \delta_{ij}$, so

$$dV = |dx^1 \wedge \cdots \wedge dx^n|,$$

and “volume” becomes ordinary volume.

There is an even more important construction associated with a Riemannian metric on M , which will occupy us for the rest of the chapter. For every C^∞ curve $\gamma: [a, b] \rightarrow M$, we have tangent vectors

$$\gamma'(t) = \frac{d\gamma}{dt} \in M_{\gamma(t)},$$

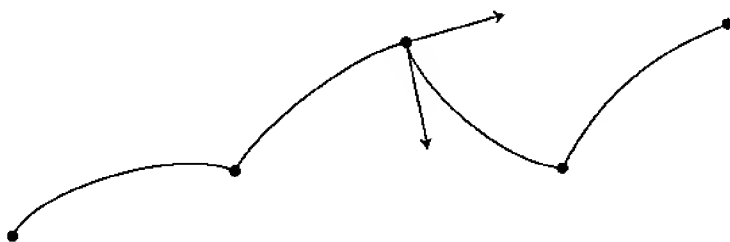
and can therefore use (\cdot, \cdot) to define their length

$$\left\| \frac{d\gamma}{dt} \right\| = \sqrt{\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle} \quad \left(= \sqrt{\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle_{\gamma(t)}}, \text{ to be precise} \right).$$

We can then define the length of γ from a to b ,

$$L_a^b(\gamma) = \int_a^b \left\| \frac{d\gamma}{dt} \right\| dt \quad \left(= \int_a^b \|\gamma'(t)\| dt \right).$$

If γ is merely *piecewise smooth*, meaning that there is a partition $a = t_0 < \cdots < t_n = b$ of $[a, b]$ such that γ is smooth on each $[t_{i-1}, t_i]$ (with possibly different



left- and right-hand derivatives at t_1, \dots, t_{n-1}), we can define the **length** of γ by

$$L_a^b(\gamma) = \sum_{i=1}^n L_{t_{i-1}}^{t_i}(\gamma|_{[t_{i-1}, t_i]}).$$

Whenever there is no possibility of misunderstanding we will denote L_a^b simply by L . A little argument shows (Problem 15) that for piecewise smooth curves in \mathbb{R}^n , with the usual Riemannian metric

$$\sum_{i=1}^n dx^i \otimes dx^i,$$

this definition agrees with the definition of length as the least upper bound of the lengths of inscribed polygonal curves.

We can also define a function $s: [a, b] \rightarrow \mathbb{R}$, the “arclength function of γ ” by

$$s(t) = L_a^t(\gamma) = \int_a^t \left\| \frac{d\gamma}{dt} \right\| dt.$$

Naturally,

$$(*) \quad s'(t) = \left\| \frac{d\gamma}{dt} \right\|.$$

Consequently $d\gamma/dt$ has constant length 1 precisely when $s(t) = t + \text{constant}$, thus precisely when $s(t) = t - a$. Then

$$b - a = s(b) = L_a^b(\gamma).$$

We can reparameterize γ to be a curve on $[0, b - a]$ by defining

$$\tilde{\gamma}(t) = \gamma(t - a).$$

For the new curve $\tilde{\gamma}$ we have

$$\begin{aligned} \text{new } s(t) &= L_0^t(\tilde{\gamma}) = L_a^{t+a}(\gamma) = \text{old } s(t + a) - \text{old } s(a) \\ &= t. \end{aligned}$$

If γ satisfies $s(t) = t$ we say that γ is **parameterized by arclength** (and then often use s instead of t to denote the argument in the domain of γ).

Classically, the norm $\| \cdot \|$ on M was denoted by ds . (This makes some sort of sense even in modern notation; equation $(*)$ says that for each curve γ and corresponding $s: [a, b] \rightarrow \mathbb{R}$ we have

$$|ds| = \gamma^*(\| \cdot \|)$$

on $[a, b]$.) Consequently, in classical books one usually sees the equation

$$ds^2 = \sum_{i,j=1}^n g_{ij} dx^i dx^j.$$

Nowadays, this is sometimes interpreted as being the equivalent of the modern equation $(\cdot, \cdot) = \sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j$, but what it always actually meant was

$$\| \cdot \|^2 = \sum_{i,j=1}^n g_{ij} dx^i dx^j.$$

The symbol $dx^i dx^j$ appearing here is *not* a classical substitute for $dx^i \otimes dx^j$ — the value $(dx^i dx^j)(p)$ of $dx^i dx^j$ at p should not be interpreted as a bilinear function at all, but as the quadratic function

$$v \mapsto dx^i(p)(v) \cdot dx^j(p)(v) \quad v \in M_p,$$

and we would use the same symbol today. The classical way of indicating $dx^i \otimes dx^j$ was very strange: one wrote

$$\sum_{i,j=1}^n g_{ij} dx^i \delta x^j \quad \text{where } dx \text{ and } \delta x \text{ are independent infinitesimals.}$$

(Classically, the Riemannian metric was not a function on tangent vectors, but the inner product of two “infinitely small displacements” dx and δx .)

Consider now a Riemannian metric (\cdot, \cdot) on a *connected* manifold M . If $p, q \in M$ are any two points, then there is at least one piecewise smooth curve $\gamma: [a, b] \rightarrow M$ from p to q (there is even a smooth curve from p to q). Define

$$d(p, q) = \inf \{ L(\gamma) : \gamma \text{ a piecewise smooth curve from } p \text{ to } q \}.$$

It is clear that $d(p, q) \geq 0$ and $d(p, p) = 0$. Moreover, if $r \in M$ is a third point, then for any $\varepsilon > 0$, we can choose piecewise smooth curves

$$\gamma_1: [a, b] \rightarrow M \text{ from } p \text{ to } q \text{ with } L(\gamma_1) - d(p, q) < \varepsilon$$

$$\gamma_2: [b, c] \rightarrow M \text{ from } q \text{ to } r \text{ with } L(\gamma_2) - d(q, r) < \varepsilon.$$

If we define $\gamma: [a, c] \rightarrow M$ to be γ_1 on $[a, b]$ and γ_2 on $[b, c]$, then γ is a piecewise smooth curve from p to r and

$$L(\gamma) = L(\gamma_1) + L(\gamma_2) < d(p, q) + d(q, r) + 2\varepsilon.$$

Since this is true for all $\varepsilon > 0$, it follows that

$$d(p, r) \leq d(p, q) + d(q, r).$$

[If we did not allow piecewise smooth curves, there would be difficulties in fitting together γ_1 and γ_2 , but d would still turn out to be the same (Problem 17).] The function $d: M \times M \rightarrow \mathbb{R}$ has all properties for a metric, except that it is not so clear that $d(p, q) > 0$ for $p \neq q$. This is made clear in the following.

7. THEOREM. The function $d: M \times M \rightarrow \mathbb{R}$ is a metric on M , and if $\rho: M \times M \rightarrow \mathbb{R}$ is the original metric on M (which makes M a manifold), then (M, d) is homeomorphic to (M, ρ) .

PROOF. Both parts of the theorem are obviously consequences of the following

7'. LEMMA. Let U be an open neighborhood of the closed ball $B = \{p \in \mathbb{R}^n : |p| \leq 1\}$, let $\langle \cdot, \cdot \rangle_e$ be the "Euclidean" or usual Riemannian metric on U ,

$$\langle \cdot, \cdot \rangle_e = \sum_{i=1}^n dx^i \otimes dx^i,$$

and let $\langle \cdot, \cdot \rangle$ be any other Riemannian metric. Let $|\cdot| = \|\cdot\|_e$ and $\|\cdot\|$ be the corresponding norms. Then there are numbers $m, M > 0$ such that

$$m \cdot |\cdot| \leq \|\cdot\| \leq M \cdot |\cdot| \quad \text{on } B,$$

and consequently for any curve $\gamma: [a, b] \rightarrow B$ we have

$$mL_e(\gamma) \leq L(\gamma) \leq ML_e(\gamma).$$

PROOF. Define $G: B \times S^{n-1} \rightarrow \mathbb{R}$ by

$$G(p, a) = \|a_p\|_p.$$

Then G is continuous and positive. Since $B \times S^{n-1}$ is compact there are numbers $m, M > 0$ such that

$$m < G < M \quad \text{on } B \times S^{n-1}.$$

Now if $p \in B$ and $0 \neq b_p \in \mathbb{R}^n_p$, let $a \in S^{n-1}$ be $a = b/|b|$. Then

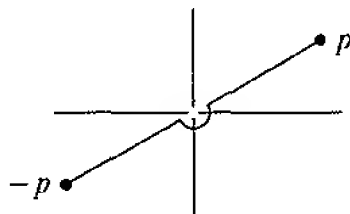
$$m|b| < |b|G(p, a) < M|b|;$$

since

$$|b|G(p, a) = |b| \cdot \|a_p\|_p = \|(|b|a)_p\|_p = \|b\|_p,$$

this gives the desired inequality (which clearly also holds for $b = 0$). ♦

Notice that the distance $d(p, q)$ defined by our metric need not be $L(\gamma)$ for any piecewise smooth curve from p to q . For example, the manifold M might be $\mathbb{R}^2 - \{0\}$, and q might be $-p$. Of course, if $d(p, q) = L(\gamma)$ for some γ ,

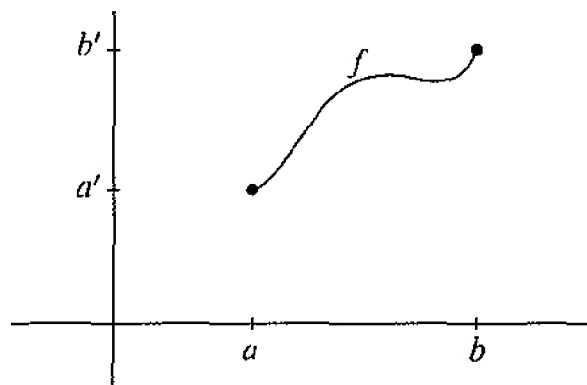


then γ is clearly a shortest piecewise smooth curve from p to q (there might be more than one shortest curve, e.g., the two semi-circles between the points p and $-p$ on S^1).

In order to investigate the question of shortest curves more thoroughly, we have to employ techniques from the “calculus of variations”. As an introduction to such techniques, we consider first a simple problem of this sort. Suppose we are given a (suitably differentiable) function

$$F: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R},$$

We seek, among all functions $f: [a, b] \rightarrow \mathbb{R}$ with $f(a) = a'$ and $f(b) = b'$ one



which will maximize (or minimize) the quantity

$$\int_a^b F(t, f(t), f'(t)) dt.$$

For example, if

$$F(t, x, y) = \sqrt{1 + y^2},$$

then we are looking for a function f on $[a, b]$ which makes the curve $t \mapsto (t, f(t))$ between (a, a') and (b, b') of shortest length

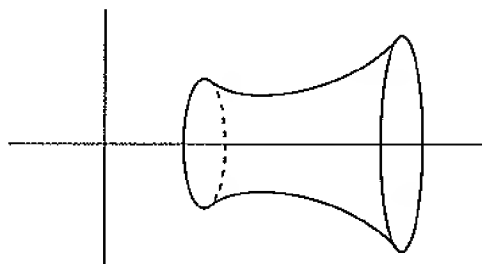
$$\int_a^b \sqrt{1 + [f'(t)]^2} dt.$$

As a second example, if

$$F(t, x, y) = 2\pi x \sqrt{1 + y^2},$$

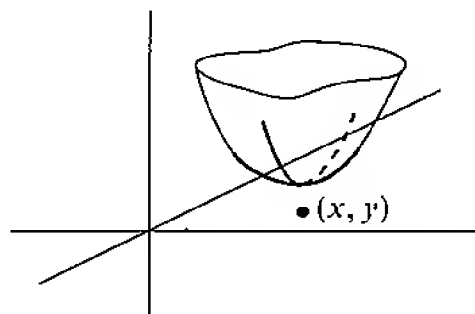
then we are trying to minimize the area of the surface obtained by revolving the graph of f around the x -axis, which is given (Problem 12) by

$$2\pi \int_a^b f(t) \sqrt{1 + [f'(t)]^2} dt.$$



To approach this sort of problem we recall first the methods used for solving the much simpler problem of determining the maximum or minimum of a function $f: \mathbb{R} \rightarrow \mathbb{R}$. To solve this problem, we examine the critical points of f , i.e., those points x for which $f'(x) = 0$. A critical point is not necessarily a maximum or minimum, or even a local maximum or minimum, but critical points are the only candidates for maxima or minima if f is everywhere differentiable. Similarly, for a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ we consider points $(x, y) \in \mathbb{R}^2$ for which

$$(*) \quad D_1 f(x, y) = D_2 f(x, y) = 0.$$



This is the same as saying that the curves

$$t \mapsto f(x + t, y)$$

$$t \mapsto f(x, y + t)$$

have derivative 0 at 0. We might try to get more information by considering the condition

$$0 = (f \circ c)'(0)$$

for every curve $c: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^2$ with $c(0) = (x, y)$, but it turns out that these conditions follow from (*), because of the chain rule.

To find maxima and minima for

$$J(f) = \int_a^b F(t, f(t), f'(t)) dt$$

we wish to proceed in an analogous way, by considering curves *in the set of all functions* $f: [a, b] \rightarrow \mathbb{R}$. This can be done by considering a “variation” of f , that is, a function

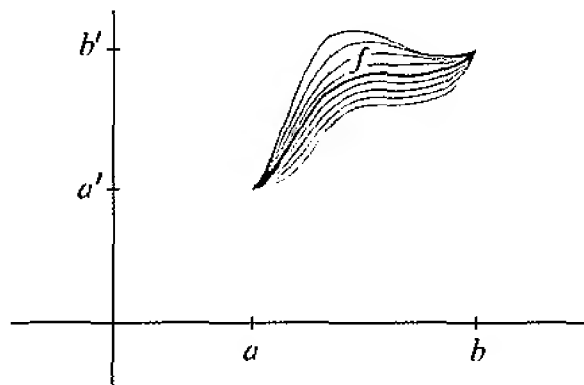
$$\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow \mathbb{R}$$

such that

$$\alpha(0, t) = f(t).$$

The functions $t \mapsto \alpha(u, t)$ are then a family of functions on $(-\varepsilon, \varepsilon)$ which pass through f for $u = 0$. We will denote this function by $\tilde{\alpha}(u)$. Thus $\tilde{\alpha}$ is a function from $(-\varepsilon, \varepsilon)$ to the set of functions $f: [a, b] \rightarrow \mathbb{R}$. If each $\tilde{\alpha}(u)$ satisfies $\tilde{\alpha}(u)(a) = a'$, $\tilde{\alpha}(u)(b) = b'$, in other words if

$$\begin{aligned}\alpha(u, a) &= a' \\ \alpha(u, b) &= b'\end{aligned}$$



for all $u \in (-\varepsilon, \varepsilon)$, then we call α a variation of f keeping endpoints fixed.

For a variation α we now compute

$$\left. \frac{dJ(\tilde{\alpha}(u))}{du} \right|_{u=0} = \left. \frac{d}{du} \right|_{u=0} \int_a^b F\left(t, \alpha(u, t), \frac{\partial \alpha}{\partial t}(u, t)\right) dt$$

$$\begin{aligned}
&= \int_a^b \left[\frac{d}{du} \Big|_{u=0} F\left(t, \alpha(u, t), \frac{\partial \alpha}{\partial t}(u, t)\right) \right] dt \\
&= \int_a^b \left[\frac{\partial \alpha}{\partial u}(0, t) \frac{\partial F}{\partial x}(t, f(t), f'(t)) \right. \\
&\quad \left. + \frac{\partial^2 \alpha}{\partial u \partial t}(0, t) \frac{\partial F}{\partial y}(t, f(t), f'(t)) \right] dt.
\end{aligned}$$

Since $\partial^2 \alpha / \partial u \partial t = \partial^2 \alpha / \partial t \partial u$, we can apply integration by parts to the second term in the integrand, thus obtaining

$$\begin{aligned}
(*) \quad \frac{dJ(\tilde{\alpha}(u))}{du} \Big|_{u=0} &= \int_a^b \frac{\partial \alpha}{\partial u}(0, t) \left[\frac{\partial F}{\partial x}(t, f(t), f'(t)) \right. \\
&\quad \left. - \frac{d}{dt} \left(\frac{\partial F}{\partial y}(t, f(t), f'(t)) \right) \right] dt \\
&\quad + \frac{\partial \alpha}{\partial u}(0, t) \frac{\partial F}{\partial y}(t, f(t), f'(t)) \Big|_a^b.
\end{aligned}$$

For variations α keeping endpoints fixed, the second term is 0, and we obtain

$$\begin{aligned}
(**) \quad \frac{dJ(\tilde{\alpha}(u))}{du} \Big|_{u=0} &= \int_a^b \frac{\partial \alpha}{\partial u}(0, t) \left[\frac{\partial F}{\partial x}(t, f(t), f'(t)) \right. \\
&\quad \left. - \frac{d}{dt} \left(\frac{\partial F}{\partial y}(t, f(t), f'(t)) \right) \right] dt.
\end{aligned}$$

In classical treatments of the calculus of variations, the variations α were taken to be of the special form

$$\alpha(u, t) = f(t) + u\eta(t),$$

for some $\eta: [a, b] \rightarrow \mathbb{R}$ with $\eta(a) = \eta(b) = 0$. Then we obtain

$$\frac{dJ(\tilde{\alpha}(u))}{du} \Big|_{u=0} = \int_a^b \eta(t) \left[\frac{\partial F}{\partial x}(t, f(t), f'(t)) - \frac{d}{dt} \left(\frac{\partial F}{\partial y}(t, f(t), f'(t)) \right) \right] dt.$$

The final result is, of course, essentially the same. The derivative $\frac{d}{du} \Big|_{u=0} J(\tilde{\alpha}(u))$ is called the “first variation” of J and is denoted classically by

$$\delta J = \int_a^b \eta \left[\frac{\partial F}{\partial x} - \frac{d}{dt} \frac{\partial F}{\partial y} \right] dt.$$

As is usual in classical notation, the arguments of functions are either put in indiscriminately or left out indiscriminately—in this case, not only are the arguments t and $(t, f(t), f'(t))$ omitted (resulting in the disappearance of the function f for which we are solving), but the dependence of δJ on α is not indicated (which can make things pretty confusing).

If f is to maximize or minimize J , then $\delta J(\alpha)$ must be 0 for every variation α of f keeping endpoints fixed. As in the case of 1-dimensional calculus, there is no reason to expect that the condition $\delta J(\alpha) = 0$ for all α will imply that f is even a local maximum or minimum for J , and we emphasize this by introducing a definition. We call f a **critical point** of J (or an **extremal** for J) if $\delta J(\alpha) = 0$ for all variations α of f keeping endpoints fixed. The particular form (**) into which we have put δJ now allows us to deduce an important condition.

8. THEOREM (EULER'S EQUATION). The C^2 function f is a critical point of J if and only if f satisfies

$$\frac{\partial F}{\partial x}(t, f(t), f'(t)) - \frac{d}{dt} \left(\frac{\partial F}{\partial y}(t, f(t), f'(t)) \right) = 0.$$

PROOF. Clearly f must make the integral in (**) vanish for *every*

$$\eta(t) = \frac{\partial \alpha}{\partial u}(0, t)$$

which vanishes at a and b . So the theorem is a consequence of the following simple

8'. LEMMA. If a continuous function $g : [a, b] \rightarrow \mathbb{R}$ satisfies

$$\int_a^b \eta(t)g(t) dt = 0$$

for every C^∞ function η on $[a, b]$ with $\eta(a) = \eta(b) = 0$, then $g = 0$.

PROOF. Choose η to be ϕg where ϕ is positive on (a, b) and $\phi(a) = \phi(b) = 0$. ♦

As an example, consider the case where $F(t, x, y) = \sqrt{1 + y^2}$. The Euler equation is

$$0 = \frac{d}{dt} \left(\frac{f'(t)}{\sqrt{1 + [f'(t)]^2}} \right).$$

so

$$0 = \frac{\sqrt{1 + f'^2} \cdot f'' - f' \cdot \frac{f''}{\sqrt{1 + f'^2}}}{(-)},$$

hence

$$0 = (1 + f'^2) f'' - f' f'' = (1 - f' + f'^2) f'',$$

which implies that $f'' = 0$, so f is linear.

Notice that we would have obtained the same result if we had considered the case $F(t, x, y) = 1 + y^2$, for then the Euler equation is simply

$$0 = \frac{d}{dt}(2f'(t)).$$

This is analogous to the situation in 1-dimensional calculus, where the critical points of \sqrt{f} are the same as those of f , since

$$(\sqrt{f})' = \frac{f'}{2\sqrt{f}}.$$

For the case of the surface of revolution, where $F(t, x, y) = x\sqrt{1 + y^2}$, the Euler equation is

$$0 = \sqrt{1 + [f'(t)]^2} - \frac{d}{dt} \left(\frac{f(t)f'(t)}{\sqrt{1 + [f'(t)]^2}} \right);$$

this leads to the equation

$$1 + f'^2 - ff'' = 0,$$

which we will also write in the classical form

$$1 + \left(\frac{dy}{dx} \right)^2 - y \frac{d^2y}{dx^2} = 0.$$

To solve this, we use one of the \aleph_0 standard tricks (leaving justification of the details to the reader). We let

$$p = y' = \frac{dy}{dx}.$$

Then

$$\frac{d^2y}{dx^2} = \frac{dp}{dx} = \frac{dp}{dy} \cdot \frac{dy}{dx} = p \frac{dp}{dy},$$

so our equation becomes

$$\begin{aligned}
 1 + p^2 - yp \frac{dp}{dy} &= 0, \\
 \frac{p}{1 + p^2} dp &= \frac{1}{y} dy, \\
 \frac{1}{2} \log(1 + p^2) &= \log y + \text{constant} \\
 y &= \text{constant} \cdot \sqrt{1 + p^2} \\
 p = \frac{dy}{dx} &= \sqrt{cy^2 - 1} \\
 \frac{dy}{\sqrt{cy^2 - 1}} &= dx
 \end{aligned}$$

and thus (see Problem 20 for the definition and properties of the “hyperbolic cosine” function \cosh and its inverse)

$$\frac{\cosh^{-1} cy}{c} = x + k.$$

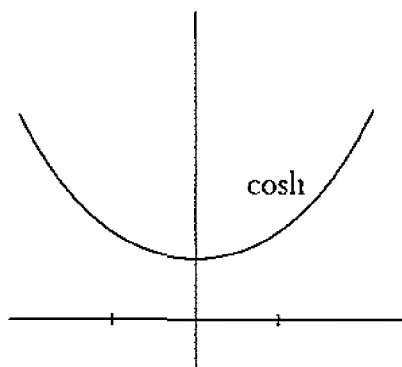
Replacing c by $1/c$, we write this as

$$(*) \quad y = c \cosh \left(\frac{x + k}{c} \right).$$

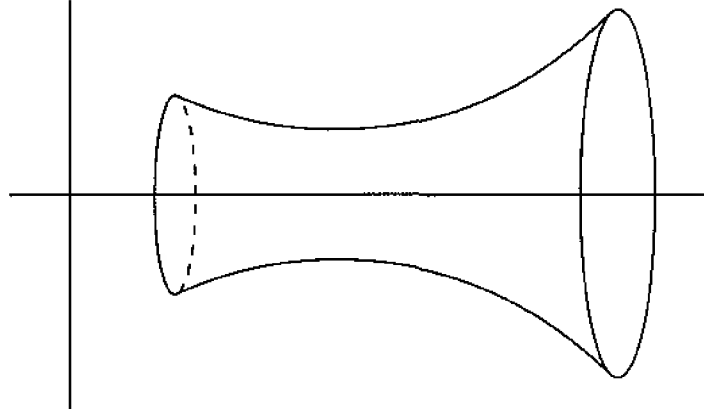
The graph of

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

is shown below; it is symmetric about the y -axis, decreasing for $x \leq 0$, and increasing for $x \geq 0$.



So our surface must look like the one drawn below. It is, by the way, not trivial to decide whether there *are* constants k and c which will make the graph of $(*)$ pass through (a, a') and (b, b') . Problem 21 investigates the special case where $a' = b'$.



It is easy to generalize these considerations to the case where $f: [a, b] \rightarrow \mathbb{R}^n$ and

$$J(f) = \int_a^b F(t, f(t), f'(t)) dt \quad \text{for } F: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}.$$

In this case we consider $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow \mathbb{R}^n$ with $\bar{\alpha}(0) = f$, and compute that

$$\begin{aligned} (***) \quad \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} &= \int_a^b \sum_{i=1}^n \frac{\partial \alpha^i}{\partial u}(0, t) \left[\frac{\partial F}{\partial x^i}(t, f(t), f'(t)) \right. \\ &\quad \left. - \frac{d}{dt} \left(\frac{\partial F}{\partial y^i}(t, f(t), f'(t)) \right) \right] dt \\ &\quad + \sum_{i=1}^n \frac{\partial \alpha^i}{\partial u}(0, t) \frac{\partial F}{\partial y^i}(t, f(t), f'(t)) \Big|_a^b. \end{aligned}$$

Thus, any critical point f of J must satisfy the n equations

$$\frac{\partial F}{\partial x^i}(t, f(t), f'(t)) - \frac{d}{dt} \left(\frac{\partial F}{\partial y^i}(t, f(t), f'(t)) \right) = 0.$$

We are now going to apply these results to the problem of finding shortest paths in a manifold M . If $\gamma: [a, b] \rightarrow M$ is a piecewise smooth curve, with $\gamma(a) = p$ and $\gamma(b) = q$, we define a variation of γ to be a function

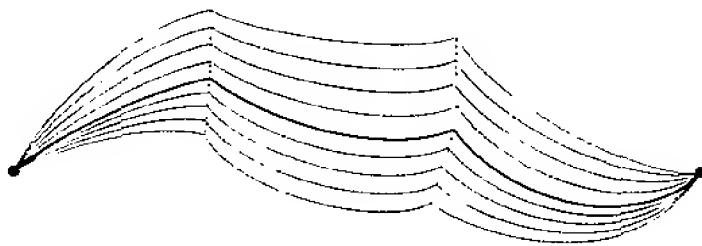
$$\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$$

for some $\varepsilon > 0$, such that

- (1) $\alpha(0, t) = \gamma(t)$,
- (2) there is a partition $a = t_0 < t_1 < \cdots < t_N = b$ of $[a, b]$ so that α is C^∞ on each strip $(-\varepsilon, \varepsilon) \times [t_{i-1}, t_i]$.

We call α a variation of γ **keeping endpoints fixed** if

$$(3) \quad \begin{aligned} \alpha(u, a) &= p \\ \alpha(u, b) &= q \end{aligned} \quad \text{for all } u \in (-\varepsilon, \varepsilon).$$



As before, we let $\tilde{\alpha}(u)$ be the path $t \mapsto \alpha(u, t)$. We would like to find which paths γ satisfy

$$\left. \frac{dL(\tilde{\alpha}(u))}{du} \right|_{u=0} = 0$$

for all variations α keeping endpoints fixed. However, we will take a hint from our first example and first find the critical points for the “energy”

$$E(\gamma) = \frac{1}{2} \int_a^b \left\| \frac{d\gamma}{dt} \right\|^2 dt = \frac{1}{2} \int_a^b \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle dt,$$

which has a much nicer integrand; afterwards we will consider the relation between the two integrals.

We can assume that each $\gamma|_{[t_{i-1}, t_i]}$ lies in some coordinate system (x, U) (otherwise we just refine the partition). If (u, t) is the standard coordinate system in $(-\varepsilon, \varepsilon) \times [a, b]$ we write

$$\begin{aligned} \frac{\partial \alpha}{\partial u}(u, t) &= \alpha_* \left(\frac{\partial}{\partial u} \Big|_{(u, t)} \right) \\ \frac{\partial \alpha}{\partial t}(u, t) &= \alpha_* \left(\frac{\partial}{\partial t} \Big|_{(u, t)} \right). \end{aligned}$$

Then $\partial\alpha/\partial t(u, t)$ is the tangent vector at time t to the curve $\tilde{\alpha}(u)$. If we adopt the abbreviations

$$\alpha^i(u, t) = x^i(\alpha(u, t)), \quad \gamma^i(t) = x^i(\gamma(t)) = \alpha^i(0, t),$$

then

$$\frac{\partial\alpha}{\partial t}(u, t) = \sum_{i=1}^n \frac{\partial\alpha^i}{\partial t}(u, t) \cdot \frac{\partial}{\partial x^i} \Big|_{\alpha(u, t)}, \quad \frac{d\gamma}{dt} = \sum_{i=1}^n \frac{d\gamma^i}{dt} \cdot \frac{\partial}{\partial x^i} \Big|_{\gamma(t)}.$$

So

$$\begin{aligned} E(\gamma | [t_{i-1}, t_i]) &= \frac{1}{2} \int_{t_{i-1}}^{t_i} \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle dt \\ &= \frac{1}{2} \int_{t_{i-1}}^{t_i} \sum_{i,j=1}^n g_{ij}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} dt. \end{aligned}$$

If we use the coordinate system x to identify U with \mathbb{R}^n , and consider the g_{ij} as functions on \mathbb{R}^n , then we are considering

$$\int_{t_{i-1}}^{t_i} F(\gamma(t), \gamma'(t)) dt$$

where

$$F(x, y) = \frac{1}{2} \sum_{i,j=1}^n g_{ij}(x) \cdot y^i y^j.$$

Then

$$\frac{\partial F}{\partial x^i} \left(\gamma(t), \frac{d\gamma}{dt} \right) = \frac{1}{2} \sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^i}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt},$$

and

$$\frac{\partial F}{\partial y^r} \left(\gamma(t), \frac{d\gamma}{dt} \right) = \sum_{r=1}^n g_{tr}(\gamma(t)) \frac{d\gamma^r}{dt},$$

so

$$\frac{d}{dt} \left(\frac{\partial F}{\partial y^r} \left(\gamma(t), \frac{d\gamma}{dt} \right) \right) = \sum_{r=1}^n g_{tr}(\gamma(t)) \frac{d^2\gamma^r}{dt^2} + \sum_{r,j=1}^n \frac{\partial g_{tr}}{\partial x^j}(\gamma(t)) \frac{d\gamma^j}{dt} \frac{d\gamma^r}{dt}.$$

In order to obtain a symmetrical looking result, we note that a little index juggling gives

$$\sum_{r,j=1}^n \frac{\partial g_{tr}}{\partial x^j} \frac{d\gamma^j}{dt} \frac{d\gamma^r}{dt} = \sum_{i,j=1}^n \frac{\partial g_{it}}{\partial x^j} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = \sum_{i,j=1}^n \frac{\partial g_{ji}}{\partial x^i} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt}$$

so

$$\sum_{r,j=1}^n \frac{\partial g_{tr}}{\partial x^j} \frac{d\gamma^j}{dt} \frac{d\gamma^r}{dt} = \frac{1}{2} \sum_{i,j=1}^n \frac{\partial g_{it}}{\partial x^j} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} + \frac{1}{2} \sum_{i,j=1}^n \frac{\partial g_{jt}}{\partial x^i} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt}.$$

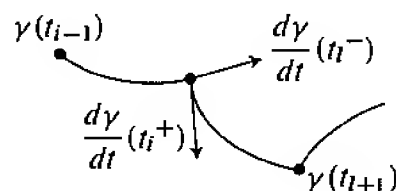
From (***) we now obtain

$$\begin{aligned} & \left. \frac{dE(\tilde{\alpha}(u) | [t_{i-1}, t_i])}{du} \right|_{u=0} \\ &= - \int_{t_{i-1}}^{t_i} \sum_{r=1}^n \frac{\partial \alpha^r}{\partial u}(0, t) \left[\sum_{r=1}^n g_{tr}(\gamma(t)) \frac{d^2 \gamma^r}{dt^2} \right. \\ & \quad \left. + \sum_{i,j=1}^n \frac{1}{2} \left(\frac{\partial g_{it}}{\partial x^j}(\gamma(t)) + \frac{\partial g_{jt}}{\partial x^i}(\gamma(t)) - \frac{\partial g_{ij}}{\partial x^t}(\gamma(t)) \right) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right] dt \\ & \quad + \sum_{r=1}^n \frac{\partial \alpha^r}{\partial u}(0, t) \sum_{r=1}^n g_{tr}(\gamma(t)) \frac{d\gamma^r}{dt} \Big|_{t_{i-1}}^{t_i}. \end{aligned}$$

Remember that γ is only piecewise C^∞ . Let

$\frac{d\gamma}{dt}(t_i^+) =$ right hand tangent vector of γ at t_i

$\frac{d\gamma}{dt}(t_i^-) =$ left hand tangent vector of γ at t_i .



Notice that the final sum in the above formula is simply

$$\left\langle \frac{\partial \alpha}{\partial u}(0, t_i), \frac{d\gamma}{dt}(t_i^-) \right\rangle - \left\langle \frac{\partial \alpha}{\partial u}(0, t_{i-1}), \frac{d\gamma}{dt}(t_{i-1}^+) \right\rangle.$$

To abbreviate the integral somewhat we introduce the symbols

$$[ij, t] = \frac{1}{2} \left(\frac{\partial g_{it}}{\partial x^j} + \frac{\partial g_{jt}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^t} \right).$$

These depend on the coordinate system, but the integral

$$- \int_{t_{i-1}}^{t_i} \sum_{l=1}^n \frac{\partial \alpha^l}{\partial u}(0, t) \left[\sum_{r=1}^n g_{lr}(\gamma(t)) \frac{d^2 \gamma^r}{dt^2} + \sum_{i,j=1}^n [ij, l](\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right] dt,$$

which appears in our result, clearly cannot. Consequently, we will use the exact same expression for each $[t_{i-1}, t_i]$, even though different coordinate systems may actually be involved (and hence different g_{ij} and γ^i).

Now we just have to add up these results. Let

$$\begin{aligned} \Delta_{t_i} \frac{d\gamma}{dt} &= \frac{d\gamma}{dt}(t_i^+) - \frac{d\gamma}{dt}(t_i^-) \quad i = 1, \dots, N-1 \\ \Delta_{t_0} \frac{d\gamma}{dt} &= \frac{d\gamma}{dt}(t_0^+) \\ \Delta_{t_N} \frac{d\gamma}{dt} &= -\frac{d\gamma}{dt}(t_N^-). \end{aligned}$$

Then we obtain the following formula (where there is a convention being used in the integral).

9. THEOREM (FIRST VARIATION FORMULA). For any variation α , we have

$$\begin{aligned} & \frac{dE(\tilde{\alpha}(u))}{du} \Big|_{u=0} \\ &= - \int_a^b \sum_{l=1}^n \frac{\partial \alpha^l}{\partial u}(0, t) \left[\sum_{r=1}^n g_{lr}(\gamma(t)) \frac{d^2 \gamma^r}{dt^2} + \sum_{i,j=1}^n [ij, l](\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right] dt \\ & \quad - \sum_{i=0}^N \left\langle \frac{\partial \alpha}{\partial u}(0, t_i), \Delta_{t_i} \frac{d\gamma}{dt} \right\rangle. \end{aligned}$$

(In the case of a variation α leaving endpoints fixed, the sum can be written from 1 to $N-1$.)

This result is not very pretty, but there it is. It should be noted that $[ij, l]$ are *not* the components of a tensor. Nevertheless, later on we will have an invariant interpretation of the first variation formula. For the time being we present, with apologies, this coordinate dependent approach. From the first variation formula it is, of course, simple to obtain conditions for critical points of E .

10. COROLLARY. If $\gamma: [a, b] \rightarrow M$ is a C^∞ path, then γ is a critical point of E_a^b if and only if for every coordinate system (x, U) we have

$$\sum_{r=1}^n g_{lr}(\gamma(t)) \frac{d^2 \gamma^r}{dt^2} + \sum_{i,j=1}^n [ij, l](\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0 \quad \text{for } \gamma(t) \in U.$$

PROOF. Suppose γ is a critical point. Given t with $\gamma(t) \in U$, choose a partition of $[a, b]$ with $t \in (t_{i-1}, t_i)$ for some i , and such that $\gamma|_{[t_{i-1}, t_i]}$ is in U . If α is a variation of γ keeping endpoints fixed, then in the first variation formula we can assume that the part of the integral from t_{i-1} to t_i is written in terms of (x, U) . The final term in the formula vanishes since γ is C^∞ . Now apply the method of proof in Lemma 8', choosing all $\partial \alpha^l / \partial u(0, t)$ to be 0, except one, which is 0 outside of (t_{i-1}, t_i) , but a positive function times the term in brackets on (t_{i-1}, t_i) . ♦

In order to put the equations of Corollary 10 in a standard form we introduce another set of symbols

$$\Gamma_{ij}^k = \sum_{l=1}^n g^{kl} [ij, l] = \sum_{l=1}^n g^{kl} \frac{1}{2} \left(\frac{\partial g_{il}}{\partial x^j} + \frac{\partial g_{jl}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l} \right).$$

Our equations can now be written

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0.$$

We know from the standard theorem about systems of second order differential equations (Problem 5-4), that for each $p \in M$ and each $v \in M_p$, there is a unique $\gamma: (-\varepsilon, \varepsilon) \rightarrow M$, for some $\varepsilon > 0$, such that γ satisfies

$$\gamma(0) = p$$

$$\frac{d\gamma}{dt}(0) = v$$

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0.$$

Moreover, this γ is C^∞ on $(-\varepsilon, \varepsilon)$. This last fact shows that if $\gamma_1: [0, \varepsilon) \rightarrow M$ and $\gamma_2: (-\varepsilon, 0] \rightarrow M$ are C^∞ functions satisfying this equation, and if moreover

$$\gamma_1(0) = \gamma_2(0)$$

$$\frac{d\gamma_1}{dt}(0^+) = \frac{d\gamma_2}{dt}(0^-),$$

then γ_1 and γ_2 together give a C^∞ function on $(-\varepsilon, \varepsilon)$. Naturally, we could replace 0 by any other t . We now have the more precise result,

11. COROLLARY. A piecewise C^∞ path $\gamma: [a, b] \rightarrow M$ is a critical point for E_a^b if and only if γ is actually C^∞ on $[a, b]$ and for every coordinate system (x, U) satisfies

$$\boxed{\frac{d^2\gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0} \quad \text{for } \gamma(t) \in U.$$

PROOF. Let γ be a critical point. Choosing the same α^l as before (all α^l are 0 outside of (t_{i-1}, t_i)), we see that $\gamma|_{[t_{i-1}, t_i]}$ satisfies the equation, because the final term in the first variation formula still vanishes. Now choose α so that

$$\frac{\partial \alpha}{\partial u}(0, t_i) = \Delta_{t_i} \frac{d\gamma}{dt}, \quad i = 1, \dots, N-1.$$

We already know that the integral in the first variation formula vanishes. So we obtain

$$0 = - \sum_{i=1}^{N-1} \left\langle \Delta_{t_i} \frac{d\gamma}{dt}, \Delta_{t_i} \frac{d\gamma}{dt} \right\rangle,$$

which implies that all $\Delta_{t_i} \frac{d\gamma}{dt}$ are 0. By our previous remarks, this means that γ is actually C^∞ on all of $[a, b]$. ♦

As the simplest possible case, consider the Euclidean metric on \mathbb{R}^n ,

$$(\ , \) = \sum_{i=1}^n dx^i \otimes dx^i.$$

Here $g_{ij} = \delta_{ij}$, so all $\partial g_{ij} / \partial x^k = 0$, and $\Gamma_{ij}^k = 0$. The critical points γ for the energy function satisfy

$$\frac{d^2\gamma^k}{dt^2} = 0.$$

Thus γ lies along a straight line, so γ is a critical point for the length function as well. The situation is now quite different from the first variational problem we considered, when we considered only curves of the form $t \mapsto (t, f(t))$. Any reparameterization of γ is also a critical point for length, since length is independent of parameterization (Problem 16). This shows that there are critical points for length which definitely aren't critical points for energy, since we have just seen that for γ to be a critical point for energy, the component functions of γ must be linear, and hence γ must be parameterized proportionally to arclength. This situation always prevails.

12. THEOREM. If $\gamma: [a, b] \rightarrow M$ is a critical point for E , then γ is parameterized proportionally to arclength.

PROOF. Observe first, from the definitions, that

$$\frac{\partial g_{ij}}{\partial x^l} = [il, j] + [jl, i].$$

Now we have

$$\begin{aligned} \frac{d}{dt} \left\| \frac{d\gamma}{dt} \right\|^2 &= \frac{d}{dt} \left(\sum_{i,j=1}^n g_{ij}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right) \\ &= \sum_{i,j=1}^n \sum_{l=1}^n \frac{\partial g_{ij}}{\partial x^l}(\gamma(t)) \frac{d\gamma^l}{dt} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} + \sum_{r,j=1}^n g_{rj}(\gamma(t)) \frac{d^2\gamma^r}{dt^2} \frac{d\gamma^j}{dt} \\ &\quad + \sum_{i,r=1}^n g_{ir}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d^2\gamma^r}{dt^2}. \end{aligned}$$

Replacing $\partial g_{ij}/\partial x^l$ by the value given above, this can be written as

$$\begin{aligned} \frac{d}{dt} \left\| \frac{d\gamma}{dt} \right\|^2 &= \sum_{j=1}^n \frac{d\gamma^j}{dt} \left(\sum_{r=1}^n g_{rj}(\gamma(t)) \frac{d^2\gamma^r}{dt^2} + \sum_{i,l=1}^n [il, j](\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^l}{dt} \right) \\ &\quad + \sum_{i=1}^n \frac{d\gamma^i}{dt} \left(\sum_{r=1}^n g_{ir}(\gamma(t)) \frac{d^2\gamma^r}{dt^2} + \sum_{j,l=1}^n [jl, i](\gamma(t)) \frac{d\gamma^j}{dt} \frac{d\gamma^l}{dt} \right). \end{aligned}$$

Since γ is a critical point for E , both terms in parentheses are 0 (Corollary 10). Thus the length $\|d\gamma/dt\|$ is constant. ♦

The formula

$$(*) \quad \frac{\partial g_{ij}}{\partial x^k} = [ik, j] + [jk, i]$$

occurring in this proof will be used on several occasions later on. It will also be useful to know a formula for $\partial g^{ij}/\partial x^k$. To derive one, we first differentiate

$$\sum_{m=1}^n g_{lm} g^{mj} = \delta_l^j$$

to obtain

$$\sum_{m=1}^n g_{lm} \frac{\partial g^{mj}}{\partial y^k} = - \sum_{m=1}^n \frac{\partial g_{lm}}{\partial y^k} g^{mj}.$$

Thus we have

$$\begin{aligned} \frac{\partial g^{ij}}{\partial y^k} &= \sum_{l,m} g^{il} g_{lm} \frac{\partial g^{mj}}{\partial y^k} = - \sum_{l,m} g^{il} g^{mj} \frac{\partial g_{lm}}{\partial y^k} \\ &= - \sum_{l,m} g^{il} g^{mj} ([lk, m] + [mk, l]) \quad \text{by } (*) \\ &= - \sum_l g^{il} \Gamma_{lk}^j - \sum_m g^{mj} \Gamma_{mk}^i, \end{aligned}$$

or

$$(**) \quad \frac{\partial g^{ij}}{\partial y^k} = - \sum_{l=1}^n (g^{il} \Gamma_{lk}^j + g^{lj} \Gamma_{lk}^i).$$

We can find the equations for critical points of the length function L in exactly the same way as we treated the energy function. For the moment we consider only paths $\gamma: [a, b] \rightarrow M$ with $d\gamma/dt \neq 0$ everywhere. For the portion $\gamma|_{[t_{i-1}, t_i]}$ of γ contained in a coordinate system (x, U) , we have

$$L(\gamma|_{[t_{i-1}, t_i]}) = \int_{t_{i-1}}^{t_i} \sqrt{\sum_{i,j=1}^n g_{ij}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt}} dt.$$

Considering our coordinate system as \mathbb{R}^n , we are now dealing with the case

$$F(x, y) = \sqrt{\sum_{i,j=1}^n g_{ij}(x) y^i y^j}.$$

We introduce the arclength function

$$s(t) = L_a^t(\gamma).$$

Then

$$\frac{ds}{dt} = \left\| \frac{d\gamma}{dt} \right\| = F\left(\gamma(t), \frac{d\gamma}{dt}\right).$$

So we have

$$\begin{aligned} \frac{\partial F}{\partial x^I}\left(\gamma(t), \frac{d\gamma}{dt}\right) &= \frac{1}{2} \frac{\sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^I}(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt}}{\frac{ds}{dt}} \\ \frac{\partial F}{\partial y^I}\left(\gamma(t), \frac{d\gamma}{dt}\right) &= \frac{\sum_{r=1}^n g_{Ir}(\gamma(t)) \frac{d\gamma^r}{dt}}{\frac{ds}{dt}}. \end{aligned}$$

After a little more calculation we finally obtain the equations for a critical point of L :

$$\frac{d^2\gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} - \frac{d\gamma^k}{dt} \frac{\frac{d^2s}{dt^2}}{\frac{ds}{dt}} = 0.$$

It is clear from this that critical points of E are also critical points of L (since they satisfy $d^2s/dt^2 = 0$). Conversely, given a critical point γ for L with $d\gamma/dt \neq 0$ everywhere, the function

$$s: [a, b] \rightarrow [0, L_a^b(\gamma)]$$

is a diffeomorphism, and we can consider the reparameterized curve

$$\gamma \circ s^{-1}: [0, L_a^b(\gamma)] \rightarrow M.$$

This reparameterized curve is automatically also a critical point for L , so it must satisfy the same differential equation. Since it is now parameterized by arclength, the third term vanishes, so $\gamma \circ s^{-1}$ is a critical point for E .

There is only one detail which remains unsettled. Conceivably a critical point for L might have a kink, but be C^∞ because it has a zero tangent vector

there, as in the figure below. In this case it would not be possible to reparam-



terize γ by arclength. Problem 37 shows that this situation cannot arise.

Henceforth we will call a critical point of E a **geodesic** on M (for the Riemannian metric (\cdot, \cdot)). This name comes from the science of geodesy, which is concerned with the measurement of the earth's surface, including surveying and the measurement of degrees of latitude and longitude. A geodesic on the earth's surface is a segment of a great circle, which is the shortest path between two points. Before we can say whether this is true for geodesics in general, which are so far merely known to be critical points for length, we must initiate a local study of geodesics.

The most elementary properties of geodesics depend only on facts about differential equations. Observe that the equations for a geodesic,

$$\frac{d^2\gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0,$$

have an important homogeneity property: if γ is a geodesic, then $t \mapsto \gamma(ct)$ is also clearly a geodesic. This feature of the equation allows us to improve the result given by the basic existence and uniqueness theorems.

13. THEOREM. Let $p \in M$. Then there is a neighborhood U of p and a number $\varepsilon > 0$ such that for every $q \in U$ and every tangent vector $v \in M_q$ with $\|v\| < \varepsilon$ there is a unique geodesic

$$\gamma_v: (-2, 2) \rightarrow M$$

satisfying

$$\gamma_v(0) = q, \quad \frac{d\gamma_v}{dt}(0) = v.$$

PROOF. The fundamental existence and uniqueness theorem says that there is a neighborhood U of p and $\varepsilon_1, \varepsilon_2 > 0$ so that for $q \in U$ and $v \in M_q$ with $\|v\| < \varepsilon_1$ there is a unique geodesic

$$\gamma_v: (-2\varepsilon_2, 2\varepsilon_2) \rightarrow M$$

with the required initial conditions.

Choose $\varepsilon < \varepsilon_1 \varepsilon_2$: Then if $|v| < \varepsilon$ and $|t| < 2$ we have

$$\|v/\varepsilon_2\| < \varepsilon_1 \quad \text{and} \quad |\varepsilon_2 t| < 2\varepsilon_2.$$

So we can define $\gamma_v(t)$ to be $\gamma_{v/\varepsilon_2}(\varepsilon_2 t)$. ♦

If $v \in M_q$ is a vector for which there is a geodesic

$$\gamma: [0, 1] \rightarrow M$$

satisfying

$$\gamma(0) = q, \quad \frac{d\gamma}{dt}(0) = v,$$

then we define the **exponential** of v to be

$$\exp(v) = \exp_q(v) = \gamma(1).$$

(The reason for this terminology will be explained in the next chapter.) The geodesic γ can thus be described as

$$\gamma(t) = \exp_q(tv).$$

Since M_q is an n -dimensional vector space, there is a natural way to give it a C^∞ structure. If $\mathcal{O} \subset M_q$ is the set of all vectors $v \in M_q$ for which $\exp_q(v)$ is defined, then the map

$$\exp_q: \mathcal{O} \rightarrow M$$

is C^∞ , since the solutions of the differential equations for geodesics have a C^∞ flow. Identifying the tangent space $(M_q)_v$ at $v \in M_q$ with M_q itself, we have an induced map

$$(\exp_q)_{v*}: M_q \rightarrow M_{\exp_q(v)}.$$

In particular, we claim that the map

$$(\exp_q)_{0*}: M_q \rightarrow M_q \quad \text{is the identity.}$$

In fact, to obtain a curve c in the manifold M_q with $dc/dt(0) = v \in M_q = (M_q)_0$, we can let $c(t) = tv$. Then $\exp_q \circ c(t) = \exp_q(tv)$, the geodesic with tangent vector v at time 0, so

$$(\exp_q)_{0*}(v) = \left. \frac{d}{dt} \right|_{t=0} \exp_q(c(t)) = v.$$

Before proving the next result, we recall some facts about the manifold TM . If (x, U) is a coordinate system on M , then for $q \in U$ we can express every vector $v \in M_q$ uniquely as

$$v = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_q.$$

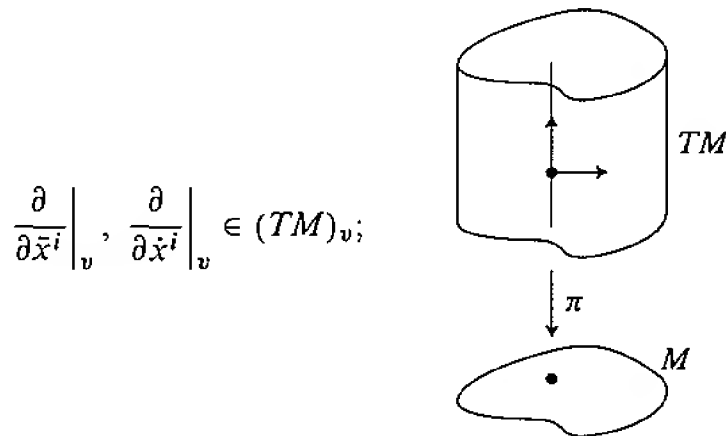
We will denote a^i by $\dot{x}^i(v)$, so that

$$v = \sum_{i=1}^n \dot{x}^i(v) \frac{\partial}{\partial x^i} \Big|_{\pi(v)},$$

where $\pi: TM \rightarrow M$ is the projection. Then

$$(x^1 \circ \pi, \dots, x^n \circ \pi, \dot{x}^1, \dots, \dot{x}^n) = (\bar{x}^1, \dots, \bar{x}^n, \dot{x}^1, \dots, \dot{x}^n)$$

is a coordinate system on $\pi^{-1}(U)$. For $v \in M_q$, $q \in U$ we therefore have tangent vectors



the vectors $\partial/\partial \dot{x}^i|_v$ are all in the tangent space of the submanifold $M_q \subset TM$, while the vectors $\partial/\partial \bar{x}^i|_v$ span a complimentary subspace.

14. THEOREM. For every $p \in M$ there is a neighborhood W and a number $\varepsilon > 0$ such that

- (1) Any two points of W are joined by a unique geodesic in M of length $< \varepsilon$.
- (2) Let $v(q, q')$ denote the unique vector $v \in M_q$ of length $< \varepsilon$ such that $\exp_q(v) = q'$. Then $(q, q') \mapsto v(q, q')$ is a C^∞ function from $W \times W \rightarrow TM$.
- (3) For each $q \in W$, the map \exp_q maps the open ε -ball in M_q diffeomorphically onto an open set $U_q \supset W$.

PROOF. Theorem 13 says that the vector $0 \in M_p$ has a neighborhood V in the manifold TM such that \exp is defined on V . Define the C^∞ function $F: V \rightarrow M \times M$ by

$$F(v) = (\pi(v), \exp(v)).$$

Let (x, U) be a coordinate system around p . We will use the coordinate system

$$(\bar{x}^1, \dots, \bar{x}^n, \dot{x}^1, \dots, \dot{x}^n),$$

described above, for $\pi^{-1}(U)$. If $\pi_i: M \times M \rightarrow M$ is projection on the i^{th} factor, then

$$(x^1 \circ \pi_1, \dots, x^n \circ \pi_1, x^1 \circ \pi_2, \dots, x^n \circ \pi_2) = (x_1^1, \dots, x_1^n, x_2^1, \dots, x_2^n)$$

is a coordinate system on $U \times U$. Now, using the fact that

$$(\exp_p)_{0*}: M_p \rightarrow M_p$$

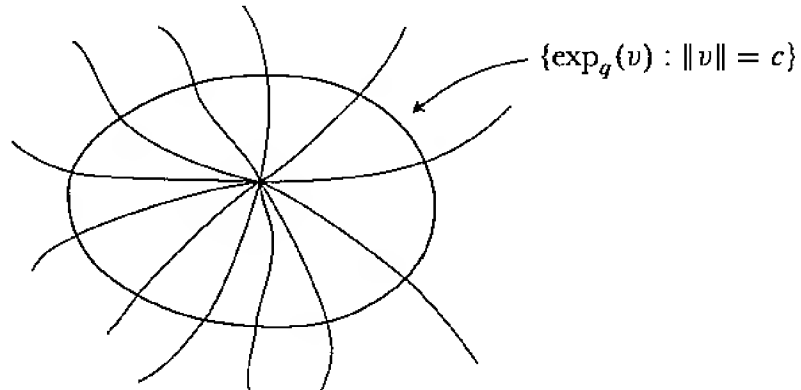
is the identity, it is not hard to see that at $0 \in M_p$ we have

$$F_* \left(\frac{\partial}{\partial \bar{x}^i} \Big|_0 \right) = \frac{\partial}{\partial x_1^i} \Big|_{(p,p)} + \frac{\partial}{\partial x_2^i} \Big|_{(p,p)}$$

$$F_* \left(\frac{\partial}{\partial \dot{x}^i} \Big|_0 \right) = \frac{\partial}{\partial x_2^i} \Big|_{(p,p)}.$$

Consequently, F_* is one-one at $0 \in M_p$, so F maps some neighborhood V' of 0 diffeomorphically onto some neighborhood of $(p, p) \in M \times M$. We may assume that V' consists of all vectors $v \in M_q$ with q in some neighborhood U' of p and $\|v\| < \varepsilon$. Choose W to be a smaller neighborhood of p for which $F(V') \supset W \times W$. ♦

Given a W as in the theorem, and $q \in W$, consider the geodesics through q of the form $t \mapsto \exp_q(tv)$ for $\|v\| < \varepsilon$. These fill out U_q . The close analysis of geodesics depends on the following.



15. LEMMA (GAUSS' LEMMA). In U_q , the geodesics through q are perpendicular to the hypersurfaces

$$\{\exp_q(v) : \|v\| = \text{constant} < \varepsilon\}.$$

FIRST PROOF. Let $v: \mathbb{R} \rightarrow M_q$ be a smooth curve with $\|v(t)\| = k$ a constant $k < \varepsilon$ for all t , and define

$$\alpha(u, t) = \exp_q(u \cdot v(t)) \quad -1 < u < 1.$$

We are claiming that for every such α we have

$$\left\langle \frac{\partial \alpha}{\partial u}(u, t), \frac{\partial \alpha}{\partial t}(u, t) \right\rangle = 0 \quad \text{for all } (u, t).$$

A calculation precisely like that in the proof of Theorem 12 proves the following equation, in which the arguments (u, t) and $\alpha(u, t)$ are omitted, for convenience:

$$(1) \quad \frac{\partial}{\partial u} \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle = \sum_{j=1}^n \frac{\partial \alpha^j}{\partial t} \left(\sum_{r=1}^n g_{rj} \frac{\partial^2 \alpha^r}{\partial u^2} + \sum_{i,l=1}^n [il, j] \frac{\partial \alpha^i}{\partial u} \frac{\partial \alpha^l}{\partial u} \right) \\ + \sum_{i=1}^n \frac{\partial \alpha^i}{\partial u} \left(\sum_{r=1}^n g_{ir} \frac{\partial^2 \alpha^r}{\partial u \partial t} + \sum_{j,l=1}^n [jl, i] \frac{\partial \alpha^j}{\partial u} \frac{\partial \alpha^l}{\partial t} \right).$$

The first term on the right is 0 since each curve $u \mapsto \alpha(u, t)$ is a geodesic. Similarly, we obtain

$$(2) \quad \frac{\partial}{\partial t} \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial u} \right\rangle = 2 \sum_{i=1}^n \frac{\partial \alpha^i}{\partial u} \left(\sum_{r=1}^n g_{ir} \frac{\partial^2 \alpha^r}{\partial u \partial t} + \sum_{j,l=1}^n [jl, i] \frac{\partial \alpha^j}{\partial u} \frac{\partial \alpha^l}{\partial t} \right),$$

which is just twice the second term on the right of (1). But $\partial \alpha / \partial u(u, t)$ is just the tangent vector at time u to the geodesic $u \mapsto \exp_q(u \cdot v(t))$, where $\|v(t)\| = k$; so $\|\partial \alpha / \partial u\| = k$. Thus the second term on the right of (2) is also 0. So

$$\left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle \quad \text{is independent of } u.$$

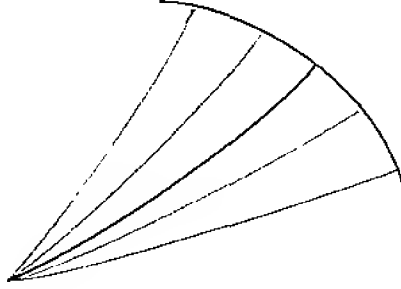
But $\alpha(0, t) = \exp_q(0) = q$, so $\partial \alpha / \partial t(0, t) = 0$. It follows that

$$\left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle = 0 \quad \text{for all } (u, t).$$

SECOND PROOF. Let $v: \mathbb{R} \rightarrow M_q$ be any smooth curve with $\|v(t)\| =$ a constant $k < \varepsilon$ for all t , and define

$$\beta(u, t) = \exp_q(t \cdot v(u)) \quad (\text{note carefully the roles played by } t \text{ and } u).$$

Then β is a variation of the geodesic $\gamma(t) = \exp_q(t \cdot v(0))$, defined on $[0, 1]$. By



the first variation formula, we have

$$\begin{aligned} \left. \frac{dE(\tilde{\beta}(u))}{du} \right|_{u=0} &= - \left\langle \frac{\partial \beta}{\partial u}(0, 1), \frac{d\gamma}{dt}(1) \right\rangle - \left\langle \frac{\partial \beta}{\partial u}(0, 0), \frac{d\gamma}{dt}(0) \right\rangle \\ &= - \left\langle \frac{\partial \beta}{\partial u}(0, 1), \frac{d\gamma}{dt}(1) \right\rangle, \end{aligned}$$

the integral vanishing since γ is a geodesic. But each curve $\tilde{\beta}(u)$ has energy

$$E(\tilde{\beta}(u)) = \int_0^1 \left\| \frac{d\tilde{\beta}(u)(t)}{dt} \right\|^2 dt = \int_0^1 k^2 dt = k^2,$$

so

$$0 = \left. \frac{dE(\tilde{\beta}(u))}{du} \right|_{u=0} = - \left\langle \frac{\partial \beta}{\partial u}(0, 1), \frac{d\gamma}{dt}(1) \right\rangle. \quad \spadesuit$$

16. COROLLARY. Let $c: [a, b] \rightarrow U_q - \{q\}$ be a piecewise smooth curve,

$$c(t) = \exp_q(u(t) \cdot v(t)),$$



for $0 < u(t) < \varepsilon$ and $\|v(t)\| = 1$. Then

$$L_a^b c \geq |u(b) - u(a)|,$$

with equality if and only if u is monotonic and v is constant, so that c is a radial geodesic joining two concentric spherical shells around q .

PROOF. If $\alpha(u, t) = \exp_q(u \cdot v(t))$, then $c(t) = \alpha(u(t), t)$ and

$$\frac{dc}{dt} = \frac{\partial \alpha}{\partial u} u'(t) + \frac{\partial \alpha}{\partial t}.$$

Since

$$\left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle = 0, \quad \left\| \frac{\partial \alpha}{\partial u} \right\| = 1,$$

we have

$$\left\| \frac{dc}{dt} \right\|^2 = |u'(t)|^2 + \left\| \frac{\partial \alpha}{\partial t} \right\|^2 \geq |u'(t)|^2,$$

with equality if and only if $\partial \alpha / \partial t = 0$, and hence $v'(t) = 0$. Thus

$$\int_0^b \left\| \frac{dc}{dt} \right\| dt \geq \int_0^b |u'(t)| dt \geq |u(b) - u(a)|,$$

with equality if and only if u is monotonic and v is constant. ♦

17. COROLLARY. Let W and ε be as in Theorem 15, let $\gamma: [0, 1] \rightarrow M$ be the geodesic of length $< \varepsilon$ joining $q, q' \in W$, and let $c: [0, 1] \rightarrow M$ be any piecewise C^∞ path from q to q' . Then

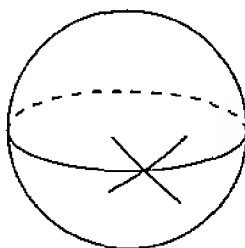
$$L(\gamma) \leq L(c),$$

with equality holding if and only if c is a reparameterization of γ .

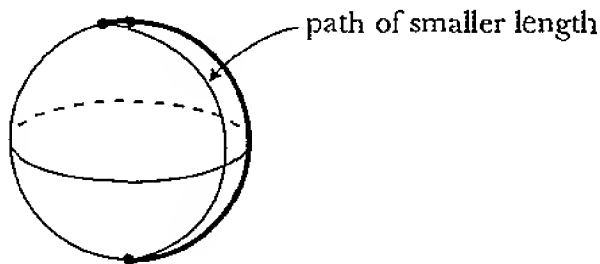
PROOF. We can assume that $q' = \exp_q(rv) \in U_q - \{q\}$ (otherwise break c up into smaller pieces). For $\delta > 0$, the path c must contain a segment which joins the spherical shell of radius δ to the spherical shell of radius r , and lies between them. By Corollary 16, the length of this segment has length $\geq r - \delta$. So the length of c is $\geq r$, and clearly c must be a reparameterization of γ for equality to hold. ♦

We thus see that *sufficiently small pieces* of geodesics are minimal paths for arc-length. We can use Corollary 17 to determine the geodesics on a few simple surfaces, without any computations, if we first introduce a notion which will play a crucial role later. If $(M, (\cdot, \cdot))$ and $(M', (\cdot, \cdot)')$ are C^∞ manifolds with Riemannian metrics, then a one-one C^∞ function $f: M \rightarrow M'$ is called an **isometry** of M into M' if $f^*(\cdot, \cdot)' = (\cdot, \cdot)$. For example, reflection through a plane $E^2 \subset \mathbb{R}^{n+1}$ is an isometry $I: S^n \rightarrow S^n$. It is clear that if $c: [0, 1] \rightarrow M$ is a C^∞ curve, then the length of c with respect to (\cdot, \cdot) is the length of $f \circ c$ with respect to $(\cdot, \cdot)'$; and if c is a geodesic, then $f \circ c$ is likewise a geodesic.

For the isometry $I: S^n \rightarrow S^n$ mentioned above, the fixed point set is the great circle $C = S^n \cap E^2$. Let $p, q \in C$ be two points with a unique geodesic C' of minimal length between them. Then $I(C')$ is a geodesic of the same length as C' between $I(p) = p$ and $I(q) = q$. So $C' = I(C')$, which implies that $C' \subset C$, so that C is a geodesic. Since there is a great circle through any point of S^n in any given direction, these are all the geodesics.

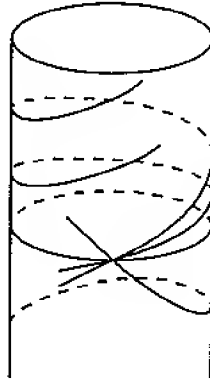


Notice that a portion of a great circle which is larger than a semi-circle is definitely not of minimal length, *even among nearby paths*. Antipodal points on

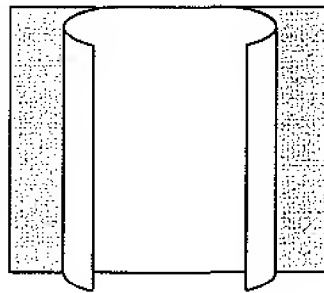
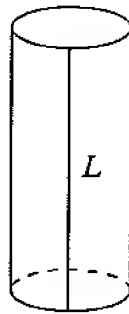


the sphere have a continuum of geodesics of minimal length between them. All other pairs of points have a unique geodesic of minimal length between them but an infinite family of non-minimal geodesics, depending on how many times the geodesic goes around the sphere and in which direction it starts.

The geodesics on a right circular cylinder Z are the generating lines, the



circles cut by planes perpendicular to the generating lines, and the helices on Z . In fact, if L is a generating line of Z , then we can set up an isometry $I: Z - L \rightarrow \mathbb{R}^2$ by rolling Z onto \mathbb{R}^2 . The geodesics on Z are just the images



under I^{-1} of the straight lines in \mathbb{R}^2 . Two points on Z have infinitely many geodesics between them.

We are now in a position to wind up our discussion of Riemannian metrics on M by establishing an important connection between the Riemannian metric (\cdot, \cdot) and the metric $d: M \times M \rightarrow \mathbb{R}$ it determines,

$$d(p, q) = \inf \{L(\gamma) : \gamma \text{ a piecewise smooth curve from } p \text{ to } q\}.$$

Notice that on both the sphere and the infinite cylinder every geodesic γ defined on an interval $[a, b]$ can be extended to a geodesic defined on all of \mathbb{R} . This is false on a cylinder of bounded height, a bounded portion of \mathbb{R}^n , or $\mathbb{R}^n - \{0\}$. In general, a manifold M with a Riemannian metric (\cdot, \cdot) is called **geodesically complete** if every geodesic $\gamma: [a, b] \rightarrow M$ can be extended to a geodesic from \mathbb{R} to M .

18. THEOREM (HOPF-RINOW-DE RHAM). If (\cdot, \cdot) is a Riemannian metric on M , then M is geodesically complete if and only if M is complete in the metric d determined by (\cdot, \cdot) . Moreover, any two points in a geodesically complete manifold can be joined by a geodesic of minimal length.

PROOF. Suppose M is geodesically complete. Given $p, q \in M$ with $d(p, q) = r > 0$, choose U_p as in Theorem 14. Let $S \subset U_p$ be the spherical shell of radius $\delta < \varepsilon$. There is a point

$$p_0 = \exp_p \delta v, \quad \|v\| = 1$$

on S such that $d(p_0, q) \leq d(s, q)$ for all $s \in S$. We claim that

$$(*) \quad \exp_p(rv) = q;$$

this will show that the geodesic $\gamma(t) = \exp_p(tv)$ is a geodesic of minimal length between p and q . To prove this result, we will prove that

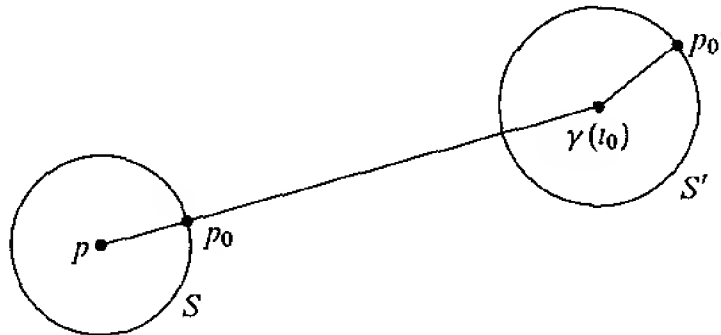
$$(**) \quad d(\gamma(t), q) = r - t \quad t \in [\delta, r].$$

First of all, since every curve from p to q must intersect S , we clearly have

$$d(p, q) = \min_{s \in S} (d(p, s) + d(s, q)) = \delta + d(p_0, q).$$

So $d(p_0, q) = r - \delta$. This proves that $(**)$ holds for $t = \delta$.

Now let $t_0 \in [\delta, r]$ be the least upper bound of all t for which $(**)$ holds. Then $(**)$ holds for t_0 also, by continuity. Suppose $t_0 < r$. Let S' be a spherical shell



of radius δ' around $\gamma(t_0)$ and let $p_0' \in S'$ be a point closest to q . Then

$$d(\gamma(t_0), q) = \min_{s \in S'} (d(\gamma(t_0), s) + d(s, q)) = \delta' + d(p_0', q),$$

so

$$(***) \quad d(p_0', q) = (r - t_0) - \delta'.$$

Hence

$$d(p, p_0') \geq d(p, q) - d(p_0', q) = t_0 + \delta'.$$

But the path c obtained by following γ from p to $\gamma(t_0)$ and then the minimal geodesic from $\gamma(t_0)$ to p_0' has length precisely $t_0 + \delta'$. So c is a path of minimal length, and must therefore be a geodesic, which means that it coincides with γ . Hence

$$\gamma(t_0 + \delta') = p_0'.$$

Hence (***) gives

$$d(\gamma(t_0 + \delta'), q) = r - (t_0 + \delta'),$$

showing that (**) holds for $t_0 + \delta'$. This contradicts the choice of t_0 , so it must be that $t_0 = r$. In other words, (**) holds for $t = r$, which proves (*).

From this result, it follows easily that M is complete with the metric d . In fact, if $A \subset M$ has diameter D , and $p \in A$, then the map $\exp_p: M_p \rightarrow M$ maps the closed disc of radius D in M_p onto a compact set containing A . In other words, bounded subsets of M have compact closure. From this it is clear that Cauchy sequences converge.

Conversely, suppose M is complete as a metric space. Given any geodesic $\gamma: (a, b) \rightarrow M$, choose $t_n \rightarrow b$. Clearly $\gamma(t_n)$ is a Cauchy sequence in M , so it converges to some point $p \in M$. Using Theorem 14, it is not difficult to show that γ can be extended past b . Consequently, by a least upper bound argument, any geodesic can be extended to \mathbb{R} . ♦

As a particular consequence of Theorem 18, note that there is always a minimal geodesic joining any two points of a compact manifold.

ADDENDUM

TUBULAR NEIGHBORHOODS

Let $M^n \subset N^{n+k}$ be a submanifold of N , with $i: M \rightarrow N$ the inclusion map, so that for every $p \in M$ we have $i_*(M_p) \subset N_p$. If (\cdot, \cdot) is a Riemannian metric for N , then we can define $M_p^\perp \subset N_p$ as

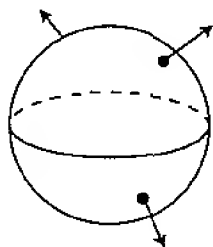
$$M_p^\perp = \{v \in N_p : (v, i_*w) = 0 \text{ for all } w \in M_p\}.$$

Let

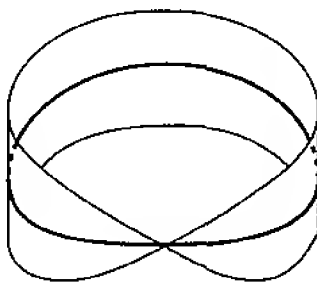
$$E = \bigcup_{p \in M} M_p^\perp \quad \text{and} \quad \varpi: E \rightarrow M \quad \text{take } M_p^\perp \text{ to } p.$$

It is not hard to see that $\nu = \varpi: E \rightarrow M$ is a k -plane bundle over M , the **normal bundle** of M in N .

For example, the normal bundle ν of $S^{n-1} \subset \mathbb{R}^n$ is the trivial 1-plane bundle, for ν has a section consisting of unit outward normal vectors. On the other hand

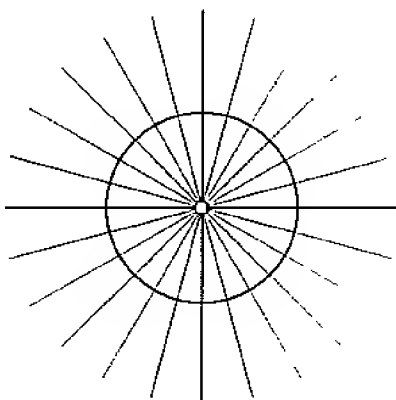


if M is the Möbius strip and $S^1 \subset M$ is a circle around the center, then it is not hard to see that the normal bundle ν will be isomorphic to the (non-trivial) bundle $M \rightarrow S^1$. If we consider $S^1 \subset M \subset \mathbb{P}^2$, then the normal bundle



of S^1 in \mathbb{P}^2 is exactly the same as the normal bundle of S^1 in M , so it too is non-trivial.

Our aim is to prove that for compact M the normal bundle of M in N is always equivalent to a bundle $\pi : U \rightarrow M$ for which U is an open neighborhood of M in N , and for which the 0-section $s : M \rightarrow U$ is just the inclusion of M into U . In the case where N is the total space of a bundle over M , this open neighborhood can be taken to be the whole total space. But in general the neighborhood cannot be all of N . For example, as an appropriate neighborhood of $S^1 \subset \mathbb{R}^2$ we can choose $\mathbb{R}^2 - \{0\}$.



A bundle $\pi : U \rightarrow M$ with U an open neighborhood of M in N , for which the 0-section $s : M \rightarrow U$ is the inclusion of M in U , is called a **tubular neighborhood** of M in N . Before proving the existence of tubular neighborhoods, we add some remarks and a Lemma.

If $\pi : U \rightarrow M$ is a tubular neighborhood, then clearly

$$\begin{aligned} \pi \circ s &= \text{identity of } M, \\ s \circ \pi &\text{ is smoothly homotopic to the identity of } U, \end{aligned}$$

so π is a deformation retraction, and $H^k(U) \approx H^k(M)$; thus M has the same de Rham cohomology as an open neighborhood. Moreover, if we choose a Riemannian metric (\cdot, \cdot) for $\pi : U \rightarrow M$ and define $D = \{e \in U : (e, e) \leq 1\}$, then D is a submanifold-with-boundary of U , and the map $\pi|_D : D \rightarrow M$ is also a deformation retraction. So M also has the same de Rham cohomology as a closed neighborhood.

19. LEMMA. Let X be a compact metric space and $X_0 \subset X$ a closed subset. Let $f : X \rightarrow Y$ be a local homeomorphism such that $f|_{X_0}$ is one-one. Then there is a neighborhood U of X_0 such that $f|_U$ is one-one.

PROOF. Let $C \subset X \times X$ be

$$\{(x, y) \in X \times X : x \neq y \text{ and } f(x) = f(y)\}.$$

Then C is closed, for if (x_n, y_n) is a sequence in C with $x_n \rightarrow x$ and $y_n \rightarrow y$, then $f(x) = \lim f(x_n) = \lim f(y_n) = f(y)$, and also $x \neq y$ since f is locally one-one.

If $g: C \rightarrow \mathbb{R}$ is $g(x, y) = d(x, X_0) + d(y, X_0)$, then $g > 0$ on C . Since C is compact, there is $\varepsilon > 0$ such that $g \geq 2\varepsilon$ on C . Then f is one-one on the ε -neighborhood of X_0 . ♦

20. THEOREM. Let $M \subset N$ be a compact submanifold of N . Then M has a tubular neighborhood $\pi: U \rightarrow M$ in N , which is equivalent to the normal bundle of M in N .

PROOF. Choose a Riemannian metric (\cdot, \cdot) for N , with the corresponding norm $\|\cdot\|$, and metric $d: N \times N \rightarrow \mathbb{R}$. Let

$$\begin{aligned} E &= \{v : v \in N_p \text{ and } v \in M_p^\perp, \text{ for some } p \in M\} \\ E_\varepsilon &= \{v \in E : \|v\| < \varepsilon\} \\ U_\varepsilon &= \{q \in N : d(q, M) < \varepsilon\}. \end{aligned}$$

It follows easily from Theorem 13, and compactness of M , that \exp is defined on E_ε for sufficiently small $\varepsilon > 0$. We claim that for sufficiently small ε , the map \exp is a diffeomorphism from E_ε onto U_ε . This will clearly prove the theorem.

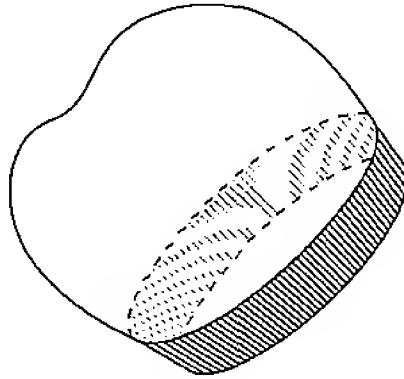
Let $V \subset E$ be the set of a non-critical points for \exp . Then $V \supset M$ (considered as a subset of E via the 0-section), and $V_1 = \overline{V \cap E_1}$ is compact; since \exp is one-one on $M \subset V_1$, it follows from Lemma 19 that for sufficiently small ε the map \exp is a diffeomorphism on E_ε .

It is clear also that $\exp(E_\varepsilon) \subset U_\varepsilon$. To prove that \exp is onto U_ε , choose any $q \in U_\varepsilon$, and a point $p \in M$ closest to q . If $\gamma: [0, 1] \rightarrow N$ is the geodesic of length $< \varepsilon$ with $\gamma(0) = p$ and $\gamma(1) = q$, it is easy to see that γ is perpendicular to M at p (compare the second proof of Gauss' Lemma). This means that $q = \exp_p d\gamma/dt(0)$ where $d\gamma/dt(0) \in E_\varepsilon$. ♦

One of the interesting features of Theorem 20 is that all the paraphernalia of Riemannian metrics and geodesics are used in its proof, while they do not even appear in the statement. Theorem 20 will be needed only in Chapter 11, where we will also need the following modification.

21. THEOREM. Let N be a manifold-with-boundary, with compact boundary ∂N . Then ∂N has (arbitrarily small) open [and closed] neighborhoods for which there are deformation retractions onto ∂N .

PROOF. Exactly the same as the proof of Theorem 20, using only inward pointing normal vectors. ♦



PROBLEMS

1. Let V be a vector space over a field F of characteristic $\neq 2$, and let $h: V \times V \rightarrow F$ be symmetric and bilinear.

(a) Define $q: V \rightarrow F$ by $q(v) = h(v, v)$. Show that if ϕ_1, \dots, ϕ_n is a basis for V^* , then

$$q = \sum_{i,j=1}^n a_{ij} v_i^* \cdot v_j^*$$

for some a_{ij} .

(b) Show that

$$q(-v) = q(v)$$

$$h(u, v) = \frac{1}{2}[q(u+v) - q(u) - q(v)].$$

(c) Suppose $q: V \rightarrow F$ satisfies $q(-v) = q(v)$, and that $h(u, v) = q(u+v) - q(u) - q(v)$ is bilinear. Show that

$$q(u+v+w) - q(u) - q(v+w) = q(u+v) - q(u) - q(v) - q(u+w) + q(u) + q(w).$$

Conclude that $q(0) = 0$, and $q(2u) = 4q(u)$. Then show that $q(v) = h(v, v)$.

2. Let (\cdot, \cdot) be a Euclidean metric for V^* . Suppose $\phi_i, \psi_i \in V^*$ satisfy $\phi_1 \wedge \dots \wedge \phi_k = \psi_1 \wedge \dots \wedge \psi_k \neq 0$, and let W_ϕ and W_ψ be the subspaces of V^* spanned by the ϕ_i and ψ_i .

(a) Show that $\omega \in W_\phi$ if and only if $\omega \wedge \phi_1 \wedge \dots \wedge \phi_k = 0$. Conclude that $W_\phi = W_\psi$.

(b) Let $\sigma_1, \dots, \sigma_k$ be an orthonormal basis of $W_\phi = W_\psi$. If $\phi_i = \sum_j a_{ji} \sigma_j$, show that the signed k -dimensional volume of the parallelepiped spanned by ϕ_1, \dots, ϕ_k is $\det(a_{ij})$. (The sign is $+$ if ϕ_1, \dots, ϕ_k has the same orientation as $\sigma_1, \dots, \sigma_k$, and $-$ otherwise.)

(c) Using Problem 7-9, show that this volume is the same for ψ_1, \dots, ψ_k .

(d) Conversely, if $W_\phi = W_\psi$, and the signed volumes of the parallelepipeds are the same, show that $\phi_1 \wedge \dots \wedge \phi_k = \psi_1 \wedge \dots \wedge \psi_k$.

If we identify V with V^{**} , so that we have a wedge product $v_1 \wedge \dots \wedge v_k$ of vectors $v_i \in V$, then we have a geometric condition for equality with $w_1 \wedge \dots \wedge w_k$. In *Leçons sur la Géométrie des Espaces de Riemann*, É. Cartan uses this condition to define $\Omega^k(V^*)$ as formal sums of equivalence classes of k vectors; he deduces geometrically the corresponding conditions on the coordinates of v_i, w_i .

3. Let V be an n -dimensional vector space, and (\cdot, \cdot) an inner product on V which is not necessarily positive definite. A basis v_1, \dots, v_n for V is called **orthonormal** if $(v_i, v_j) = \pm \delta_{ij}$.

- (a) If $V \neq \{0\}$, then there is a vector $v \in V$ with $(v, v) \neq 0$.
- (b) For $W \subset V$, let $W^\perp = \{v \in V : (v, w) = 0 \text{ for all } w \in W\}$. Prove that $\dim W^\perp \geq n - \dim W$. *Hint:* If $\{w_i\}$ is a basis for W , consider the linear functionals $\lambda_i: V \rightarrow \mathbb{R}$ defined by $\lambda_i(v) = (v, w_i)$.
- (c) If $(\ , \)$ is non-degenerate on W , then $V = W \oplus W^\perp$, and $(\ , \)$ is also non-degenerate on W^\perp .
- (d) V has an orthonormal basis. Thus, there is an isomorphism $f: \mathbb{R}^n \rightarrow V$ with $f^*(\ , \) = (\ , \)_r$ for some r (the inner product $(\ , \)_r$ is defined on page 301).
- (e) The **index** of $(\ , \)$ is the largest dimension of a subspace $W \subset V$ such that $(\ , \)|_W$ is negative definite. Show that the index is $n - r$, thus showing that r is unique ("Sylvester's Law of Inertia").
4. Let $(\ , \)$ be a (possibly non-positive definite) inner product on V , and let v_1, \dots, v_n be an orthonormal basis (see Problem 3). Define an inner product $(\ , \)^k$ on $\Omega^k(V)$ by requiring that

$$v_{i_1}^* \wedge \cdots \wedge v_{i_k}^* \quad 1 \leq i_1 < \cdots < i_k \leq n$$

be an orthonormal basis, with

$$\langle v_{i_1}^* \wedge \cdots \wedge v_{i_k}^*, v_{j_1}^* \wedge \cdots \wedge v_{j_k}^* \rangle^k = \det((v_{i_\alpha}, v_{j_\beta})).$$

- (a) Show that $(\ , \)^k$ is independent of the basis v_1, \dots, v_n . (Use Problem 7-16.)
- (b) Show that

$$(\phi_1 \wedge \cdots \wedge \phi_k, \psi_1 \wedge \cdots \wedge \psi_k)^k = \det((\phi_i, \psi_j)^*) = \det((\phi_i, \psi_j)^1).$$

- (c) If $(\ , \)$ has index i , then

$$(v_1^* \wedge \cdots \wedge v_n^*, v_1^* \wedge \cdots \wedge v_n^*)^n = (-1)^i.$$

- (d) For those who know about \otimes and Λ^k . Using the isomorphisms $\otimes^k V^* \approx (\otimes^k V)^*$ and $\Lambda^k(V^*) \approx (\Lambda^k V)^*$, define inner products on $\otimes^k V$ and $\Lambda^k V$ by using the isomorphism $V \rightarrow V^*$ given by the inner product on V . Show that these inner products agree with the ones defined above.

5. Recall the definition of $v_1 \times \cdots \times v_{n-1}$ in Problem 7-26.

- (a) Show that $\langle v_1 \times \cdots \times v_{n-1}, v_i \rangle = 0$.
- (b) Show that $|v_1 \times \cdots \times v_{n-1}| = \sqrt{\det(g_{ij})}$, where $g_{ij} = \langle v_i, v_j \rangle$. *Hint:* Apply the result on page 308 to a certain $(n-1)$ -dimensional subspace of \mathbb{R}^n .

6. Let $\xi = \pi: E \rightarrow B$ be a vector bundle. An **indefinite metric** on ξ is a continuous choice of a non-positive definite inner product $(\ , \)_p$ on each $\pi^{-1}(p)$. Show that the index of $(\ , \)_p$ is constant on each component of B .

7. This problem requires a little knowledge of simple-connectedness and covering spaces.

(a) There is no way of continuously choosing a 1-dimensional subspace of S^2_p , for each $p \in S^2$. (Consider the space consisting of the two unit vectors in each subspace.)

(b) There is no Riemannian metric of index 1 on S^2 .

8. Let $(\ , \)$ and $(\ , \)'$ be two Riemannian metrics on a vector bundle $\xi = \pi: E \rightarrow B$. Let S be the set of $e \in E$ with $(e, e) = 1$, and define S' similarly. Show that S is homeomorphic to S' . If ξ is a smooth bundle over a manifold M , show that S is diffeomorphic to S' .

9. Show by a computation that if the functions g_{ij} and g'_{ij} are related by

$$g'_{\alpha\beta} = \sum_{i,j} g_{ij} \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x^j}{\partial x'^\beta},$$

with $\det(g_{ij}) \neq 0$, and the functions g^{ij} , g'^{ij} are defined by

$$\sum_{k=1}^n g^{ik} g_{kj} = \delta_j^i, \quad \sum_{k=1}^n g'^{ik} g'_{kj} = \delta_j^i,$$

then

$$g'^{\alpha\beta} = \sum_{i,j} g^{ij} \frac{\partial x'^\alpha}{\partial x^i} \frac{\partial x'^\beta}{\partial x^j}.$$

This, of course, is the classical way of defining the tensor [having the components] g^{ij} .

10. (a) Let $(\ , \)$ be a Riemannian metric on M , and A a tensor of type $\binom{1}{1}$, so that $A(p): M_p \rightarrow M_p$. Define a tensor B of type $\binom{2}{0}$ by

$$B(p)(v_1, v_2) = (A(p)(v_1), v_2).$$

If the expression for A in a coordinate system is

$$A = \sum_{i,j=1}^n A_i^j dx^i \otimes \frac{\partial}{\partial x^j},$$

show that $B = \sum_{i,k} B_{ik} dx^i \otimes dx^k$, where

$$B_{ik} = \sum_{j=1}^n A_i^j g_{jk}.$$

(b) Similarly, define a tensor C of type $\binom{0}{2}$ by

$$C(p)(\lambda_1, \lambda_2) = (A(p))^*(\lambda_1), \lambda_2).$$

Show that if C has components C^{kj} , then

$$C^{kj} = \sum_{i=1}^n g^{ki} A_i^j.$$

The tensors B and C are said to be obtained from A by “raising and lowering indices”.

11. (a) Let X_1, \dots, X_n be linearly independent vector fields on a manifold M with a Riemannian metric (\cdot, \cdot) . Show that the Gram-Schmidt process can be applied to the vector fields all at once, so that we obtain n everywhere orthonormal vector fields Y_1, \dots, Y_n .

(b) For the case of a non-positive definite metric, find Y_1, \dots, Y_n with $(Y_i, Y_j) = \pm \delta_{ij}$ in a neighborhood of any point.

12. (a) If $f: [a, b] \rightarrow \mathbb{R}$ is positive, show that the area of the surface obtained by revolving the graph of f around the x -axis is

$$\int_a^b 2\pi f \sqrt{1 + (f')^2}.$$

(b) Compute the area of S^2 .

13. Let $M \subset \mathbb{R}^n$ be an $(n-1)$ -dimensional submanifold with orientation μ . The **outward unit normal** $\nu(p)$ at $p \in M$ is defined to be that vector in \mathbb{R}^n_p of length 1 such that $\nu(p), (v_1)_p, \dots, (v_{n-1})_p$ is positively oriented in \mathbb{R}^n_p when $(v_1)_p, \dots, (v_{n-1})_p$ is positively oriented in M_p .

(a) If $M = \partial N$ for an n -dimensional manifold-with-boundary $N \subset \mathbb{R}^n$, then $\nu(p)$ is outward pointing in the sense of Chapter 8.

(b) Let dV_{n-1} be the volume element of M determined by the Riemannian metric it acquires as a submanifold of \mathbb{R}^n . Show that if we consider $\nu(p)$ as an element of \mathbb{R}^n , then

$$dV_{n-1}(p)((v_1)_p, \dots, (v_{n-1})_p) = \det \begin{pmatrix} \nu(p) \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix}.$$

Conclude that $dV_{n-1}(p)$ is the restriction to M_p of

$$\sum_{i=1}^n (-1)^{i-1} v^i(p) dx^1(p) \wedge \cdots \wedge \widehat{dx^i(p)} \wedge \cdots \wedge dx^n(p).$$

(c) Note that $v_1 \times \cdots \times v_{n-1} = \alpha v(p)$ for some $\alpha \in \mathbb{R}$ (by Problem 5). Show that for $w \in \mathbb{R}^n$ we have

$$(w, v(p)) \cdot (v_1 \times \cdots \times v_{n-1}, v(p)) = (w, v_1 \times \cdots \times v_{n-1}).$$

Conclude that

$$v^i(p) \cdot dV_{n-1}(p) = \text{restriction to } M_p \text{ of} \\ (-1)^{i-1} dx^1(p) \wedge \cdots \wedge \widehat{dx^i(p)} \wedge \cdots \wedge dx^n(p).$$

(d) Let $M \subset \mathbb{R}^n$ be a compact n -dimensional manifold-with-boundary, with v the outward unit normal on ∂M . Denote the volume element of M by dV_n , and that of ∂M by dV_{n-1} . Let $X = \sum_i a^i \partial/\partial x^i$ be a vector field on M . Prove the *Divergence Theorem*:

$$\int_M \operatorname{div} X dV_n = \int_{\partial M} (X, v) dV_{n-1}$$

(the function $\operatorname{div} X$ is defined in Problem 7-27). *Hint*: Consider the form ω on M defined by

$$\omega = \sum_{i=1}^n (-1)^{i-1} a^i dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^n.$$

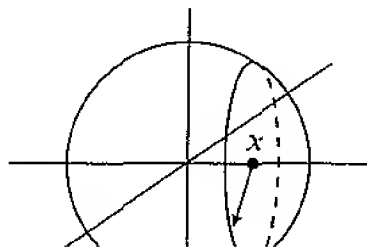
(e) Let $M \subset \mathbb{R}^3$ be a compact 2-dimensional manifold-with-boundary, with orientation μ , and outward unit normal v . Let T be the vector field on ∂M consisting of positively oriented unit vectors. Denote the volume element of M by dA , and that of ∂M by ds . Let X be a vector field on M . Prove (the original) *Stokes' Theorem*:

$$\int_M (\nabla \times X, v) dA = \int_{\partial M} (X, T) ds$$

($\nabla \times X$ is defined in Problem 7-27).

14. (a) Let V_n be the volume of the unit ball in \mathbb{R}^n . Show that

$$V_n = \int_{-1}^1 (1-x^2)^{(n-1)/2} V_{n-1} dx.$$



(b) If $I_n = \int_{-1}^1 (1 - x^2)^{(n-1)/2} dx$, show that

$$I_n = \frac{n-1}{n} I_{n-2}.$$

(c) Using $V_1 = 2$, $V_2 = \pi$, show that

$$V_n = \begin{cases} \frac{\pi^{n/2}}{(n/2)!} & n \text{ even} \\ \frac{2^{(n+1)/2} \pi^{(n-1)/2}}{1 \cdot 3 \cdot 5 \cdots n} & n \text{ odd.} \end{cases}$$

(In terms of the Γ function, this can be written $\frac{\pi^{n/2}}{\Gamma(1 + n/2)}$.)

(d) Let A_{n-1} be the $(n-1)$ -volume of S^{n-1} . Using the method of proof in Corollary 8-8, but reversing the order of integration, show that

$$V_n = \int_0^1 r^{n-1} A_{n-1} dr = \frac{1}{n} A_{n-1}.$$

(e) Obtain this same result by applying the Divergence Theorem (Problem 13), with $X(p) = p_p$.

15. (a) Let $c: [0, 1] \rightarrow \mathbb{R}^n$ be a differentiable curve, where \mathbb{R}^n has the usual Riemannian metric $(\cdot, \cdot) = \sum_i dx^i \otimes dx^i$. Show that

$$L(c) = \int_0^1 \sqrt{\sum_{i=1}^n [(c^i)'(t)]^2} dt.$$

(b) For the special case $c: [0, 1] \rightarrow \mathbb{R}^2$ given by $c(t) = (t, f(t))$, show that this length,

$$\int_0^1 \sqrt{1 + [f'(t)]^2} dt,$$

is the least upper bound of the lengths of inscribed polygonal curves.



Hint: If the inscribed polygonal curve is determined by the points $(t_i, c(t_i))$ for

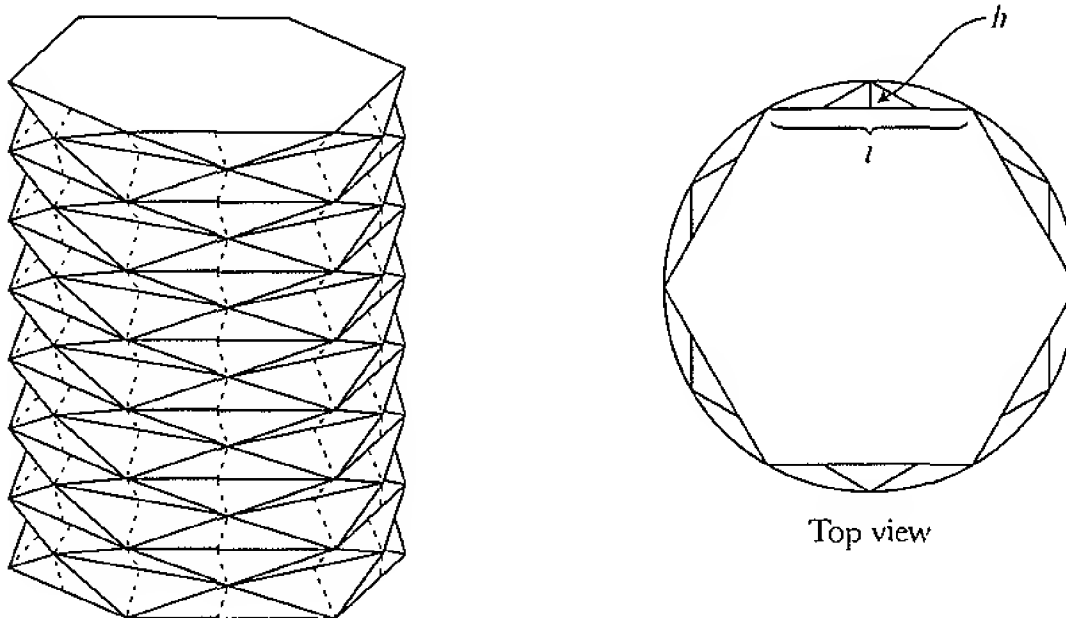
a partition $0 = t_0 < \cdots < t_n = 1$ of $[0, 1]$, then we have

$$\begin{aligned} |c(t_i) - c(t_{i-1})| &= \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2} \\ &= \sqrt{(t_i - t_{i-1})^2 + f'(\xi_i)(t_i - t_{i-1})^2} \end{aligned}$$

for some $\xi_i \in [t_{i-1}, t_i]$.

(c) Prove the same result in the general case. *Hint:* Use the results of Problem 8-1, and uniform continuity of $\sqrt{\quad}$ on a compact set.

It is natural to suppose that the area of a surface is, similarly, the least upper bound of the areas of inscribed polygonal surfaces, but as H. Schwarz first observed, this least upper bound is infinite for a bounded portion of a cylinder! To illustrate Schwarz's example I have plagiarized the following picture from a book called *Математический Анализ на Многообразиях*, written by someone called М. Спивак.



To increase the number of triangles, we maintain the hexagonal arrangement, but move the planes of the hexagons closer together, so that the triangles are more nearly in a plane parallel to the bases of the cylinder. In this way, we can increase the number of triangles indefinitely, while the area of each approaches $hl/2$.

The topic of surface area for non-differentiable surfaces is a complex one, which we will not go into here.

16. Let $c: [0, 1] \rightarrow M$ be a curve in a manifold M with a Riemannian metric (\cdot, \cdot) . If $p: [0, 1] \rightarrow [0, 1]$ is a diffeomorphism, show that

$$L(c) = L(c \circ p).$$

17. Show that the metric d on M may be defined using C^∞ , instead of piecewise C^∞ curves. (Show how to round off corners of a piecewise C^∞ path so that the length increases by less than any given $\varepsilon > 0$; remember that the formula for length involves only first derivatives.)

18. (a) Let $B \subset M$ be homeomorphic to the ball $\{p \in \mathbb{R}^n : |p| \leq 1\}$ and let $S \subset M$ be the subset corresponding to $\{p \in \mathbb{R}^n : |p| = 1\}$. Show that $M - S$ is disconnected, by showing that $M - B$ and $B - S$ are disjoint open subsets of $M - S$.

(b) If $p \in B - S$ and $q \in M - B$, show that $d(p, q) \geq \min_{q' \in S} d(p, q')$. Use this fact and Lemma 7' to complete the proof of Theorem 7. (In the theory of infinite dimensional manifolds, these details become quite important, for $M - S$ does *not* have to be disconnected, and Theorem 7 is false.)

19. (a) By applying integration by parts to the equation on pages 318–319, show that

$$\left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_a^b \frac{\partial^2 \alpha}{\partial u \partial t}(0, t) \left[\frac{\partial F}{\partial y}(t, f(t), f'(t)) - \int_a^t \frac{\partial F}{\partial x}(t, f(t), f'(t)) dt \right] dt;$$

this result makes sense even if f is only C^1 .

(b) *Du Bois Reymond's Lemma*. If a continuous function g on $[a, b]$ satisfies

$$\int_a^b \eta'(t)g(t) dt = 0$$

for all C^∞ functions η on $[a, b]$ with $\eta(a) = \eta(b) = 0$, then g is a constant.

Hint: The constant c must be

$$c = \frac{1}{b-a} \int_a^b g(u) du.$$

We clearly have

$$\int_a^b \eta'(t)[g(t) - c] dt = 0,$$

so we need to find a suitable η with $\eta'(t) = g(t) - c$.

(c) Conclude that if the C^1 function f is a critical point of J , then f still satisfies the Euler equations (which are not *a priori* meaningful if f is not C^2).

20. The hyperbolic sine, hyperbolic cosine, and hyperbolic tangent functions \sinh , \cosh , and \tanh are defined by

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}, \quad \tanh x = \frac{\sinh x}{\cosh x}.$$

(a) Graph \sinh , \cosh , and \tanh .

(b) Show that

$$\cosh^2 - \sinh^2 = 1$$

$$\tanh^2 + 1/\cosh^2 = 1$$

$$\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$$

$$\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y$$

$$\sinh' = \cosh$$

$$\cosh' = \sinh.$$

(c) For those who know about complex power series:

$$\sinh x = \frac{\sin ix}{i}, \quad \cosh x = \cos ix.$$

(d) The inverse functions of \sinh and \tanh are denoted by \sinh^{-1} and \tanh^{-1} , respectively, while \cosh^{-1} denotes the inverse of $\cosh | [0, \infty)$. Show that

$$\sinh(\cosh^{-1} x) = \sqrt{x^2 - 1}$$

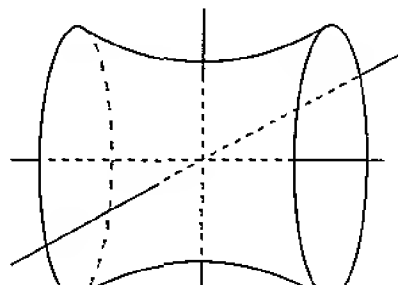
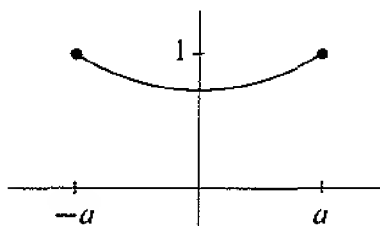
$$\cosh(\sinh^{-1} x) = \sqrt{1 + x^2}$$

$$\cosh(\tanh^{-1} x) = \frac{1}{\sqrt{1 - x^2}}$$

$$(\sinh^{-1})'(x) = \frac{1}{\sqrt{1 + x^2}}$$

$$(\cosh^{-1})'(x) = \frac{1}{\sqrt{x^2 - 1}}.$$

21. Consider the problem of finding a surface of revolution joining two circles of radius 1, situated, for convenience, at a and $-a$. We are looking for a function



of the form

$$f(x) = c \cosh \frac{x}{c}$$

where c is supposed to satisfy

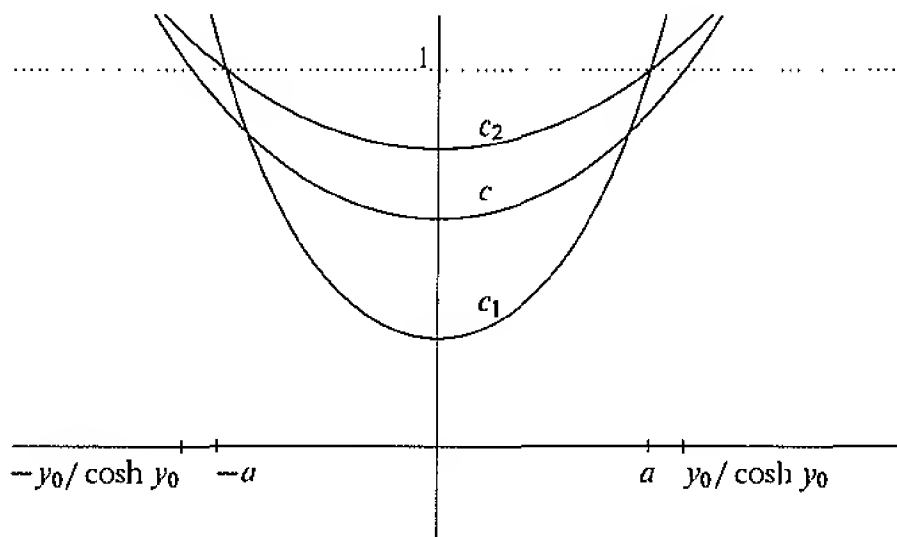
$$c \cosh \frac{a}{c} = 1 \quad (c > 0).$$

- (a) There is a unique $y_0 > 0$ with $\tanh y_0 = 1/y_0$. Examine the sign of $1/y - \tanh y$ for $y > 0$.
 (b) Examine the sign of $\cosh y - y \sinh y$ for $y > 0$.
 (c) Let

$$A_a(c) = c \cosh \frac{a}{c} \quad c > 0.$$

Show that A_a has a minimum at a/y_0 , find the value of A_a there, and sketch the graph.

- (d) There exists c with $c \cosh a/c = 1$ if and only if $a \leq y_0/\cosh y_0$. If $a = y_0/\cosh y_0$, then there is a unique such c , namely $c = a/y_0 = 1/\cosh y_0$. If $a < y_0/\cosh y_0$, then there are two such c , with $c_1 < a/y_0 < c_2$. It turns out that the surface for c_2 has smaller area.



- (e) Using Problem 20(d), show that

$$\frac{y_0}{\cosh y_0} = \sqrt{y_0^2 - 1}.$$

[$y_0 \sim 1.2$, so $\sqrt{y_0^2 - 1} \sim .67$.]

[These phenomena can be pictured more easily if we use the notion of an envelope—c.f. Volume III, pp. 176ff. The envelope of the 1-parameter family of curves

$$f_c(x) = c \cosh \frac{x}{c}$$

is determined by solving the equations

$$0 = \frac{\partial f_c(x)}{\partial c} = \cosh \frac{x}{c} - \frac{x}{c} \sinh \frac{x}{c}.$$

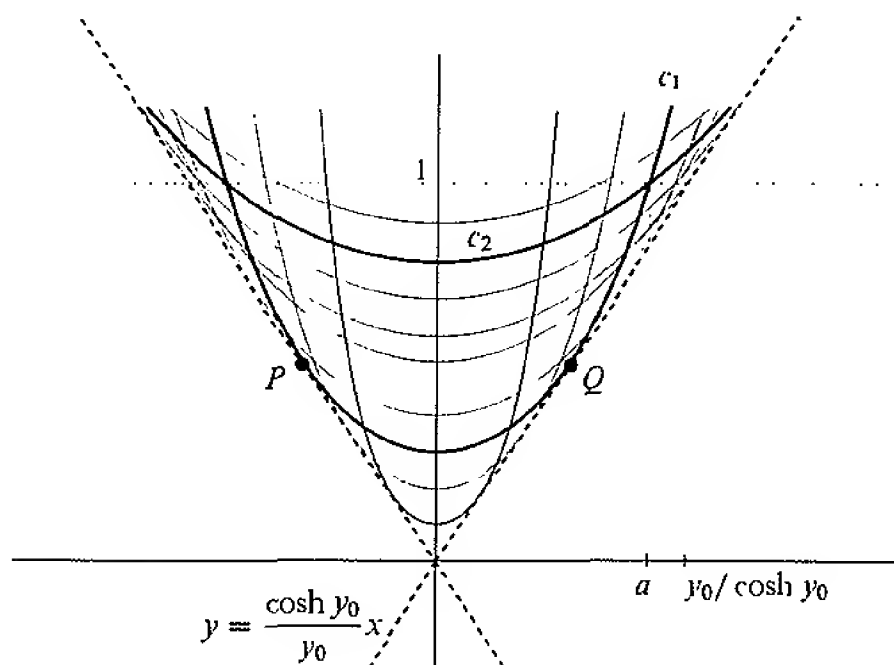
We obtain

$$\frac{x}{c} = \pm y_0, \quad y = c \cosh \frac{x}{c} = c \cosh y_0,$$

so the envelope consists of the straight lines

$$y = \pm \frac{\cosh y_0}{y_0} x.$$

The unique member of the family through $(y_0/\cosh y_0, 1)$ is tangent to the envelope at that point. For $a < y_0/\cosh y_0$, the graph of f_{c_1} is tangent to



the envelope at points $P, Q \in (-a, a)$, but the graph of f_{c_2} is tangent to the envelope at points outside $[-a, a]$. The point Q is called **conjugate** to P along the extremal f_{c_1} , and it is shown in the calculus of variations that the existence of this conjugate point implies that the portion of f_{c_1} from P to $(a, 1)$ does *not*

give a local minimum for $\int f\sqrt{1+(f')^2}$. (Compare with the discussion of conjugate points of a geodesic in Volume IV, Chapter 8, and note the remark on pg. V.396.)]

22. All of our illustrations of calculus of variations problems involved an F which does not involve t , so that the Euler equations are actually

$$\frac{\partial F}{\partial x}(f(t), f'(t)) - \frac{d}{dt} \left(\frac{\partial F}{\partial y}(f(t), f'(t)) \right) = 0.$$

(a) Show that for any f and $F: \mathbb{R}^2 \rightarrow \mathbb{R}$ we have

$$\frac{d}{dt} \left(F - f' \frac{\partial F}{\partial y} \right) = f' \left[\frac{\partial F}{\partial x} - \frac{d}{dt} \frac{\partial F}{\partial y} \right],$$

and conclude that the extremals for our problem satisfy

$$F - f' \frac{\partial F}{\partial y} = 0.$$

(b) Apply this to $F(x, y) = x\sqrt{1+y^2}$ to obtain directly the equation $dy/dx = \sqrt{c^2 y^2 - 1}$ which we eventually obtained in our solution to the problem.

23. (a) Let x and x' be two coordinate systems, with corresponding g_{ij} and g'_{ij} for the expression of a Riemannian metric. Show that

$$\begin{aligned} \frac{\partial g'_{\alpha\beta}}{\partial x'^\gamma} &= \sum_{i,j,k=1}^n \frac{\partial g_{ij}}{\partial x^k} \frac{\partial x^k}{\partial x'^\gamma} \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x^j}{\partial x'^\beta} \\ &\quad + \sum_{i,j=1}^n g_{ij} \left(\frac{\partial x^i}{\partial x'^\alpha} \frac{\partial^2 x^j}{\partial x'^\beta \partial x'^\gamma} + \frac{\partial x^j}{\partial x'^\beta} \frac{\partial^2 x^i}{\partial x'^\alpha \partial x'^\gamma} \right). \end{aligned}$$

(b) For the corresponding $[ij, k]$ and $[\alpha\beta, \gamma]'$, show that

$$[\alpha\beta, \gamma]' = \sum_{i,j,k=1}^n [ij, k] \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x^j}{\partial x'^\beta} \frac{\partial x^k}{\partial x'^\gamma} + \sum_{i,j=1}^n \frac{\partial x^i}{\partial x'^\gamma} \frac{\partial^2 x^j}{\partial x'^\alpha \partial x'^\beta},$$

so that $[ij, k]$ are not the components of a tensor.

(c) Also show that

$$\Gamma'^\gamma_{\alpha\beta} = \sum_{i,j,k=1}^n \Gamma^k_{ij} \frac{\partial x^i}{\partial x'^\alpha} \frac{\partial x^j}{\partial x'^\beta} \frac{\partial x^k}{\partial x'^\gamma} + \sum_{l=1}^n \frac{\partial^2 x^l}{\partial x'^\alpha \partial x'^\beta} \frac{\partial x'^\gamma}{\partial x^l}.$$

24. Show that any C^∞ structure on \mathbb{R} is diffeomorphic to the usual C^∞ structure. (Consider the arclength function on a geodesic for some Riemannian metric on \mathbb{R} .)

25. Let $(\ , \) = \sum_i dx^i \otimes dx^i$ be the usual Riemannian metric on \mathbb{R}^n , and let $\sum_{i,j} g_{ij} du^i \otimes du^j$ be another metric, where u^1, \dots, u^n again denotes the standard coordinate system on \mathbb{R}^n . Suppose we are told that there is a diffeomorphism $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\sum_{i,j} g_{ij} du^i \otimes du^j = f^*(\ , \)$. How can we go about finding f ?

(a) Let $\partial f / \partial u^i = e_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$. If we consider e_i as a vector field on \mathbb{R}^n , show that $f_*(\partial / \partial u^i) = e_i$.

(b) Show that $g_{ij} = (e_i, e_j)$.

(c) To solve for f it is, in theory at least, sufficient to solve for the e_i , and to solve for these we want to find differential equations

$$\frac{\partial e_i}{\partial u^j} = \sum_{r=1}^n A_{ij}^r e_r$$

satisfied by the e_i 's. Show that we must have

$$B_{ij,k} \stackrel{\text{def}}{=} \sum_{r=1}^n g_{kr} A_{ij}^r = \sum_{l=1}^n \frac{\partial^2 f^l}{\partial u^i \partial u^j} \cdot \frac{\partial f^l}{\partial u^k}.$$

(d) Show that

$$\begin{aligned} \frac{\partial g_{ij}}{\partial u^k} &= \sum_{l=1}^n \frac{\partial^2 f^l}{\partial u^i \partial u^k} \frac{\partial f^l}{\partial u^j} + \frac{\partial^2 f^l}{\partial u^j \partial u^k} \frac{\partial f^l}{\partial u^i} \\ &= \sum_{r=1}^n g_{jr} A_{ik}^r + g_{ir} A_{kj}^r \\ &= B_{ik,j} + B_{kj,i}. \end{aligned}$$

(e) By cyclically permuting i, j, k , deduce that

$$B_{ij,k} = [ij, k],$$

so that $A_{ij}^r = \Gamma_{ij}^r$. In *Leçons sur la Géométrie des Espaces de Riemann*, É. Cartan uses this approach to motivate the introduction of the Γ_{ij}^k .

(f) Deduce the result $A_{ij}^r = \Gamma_{ij}^r$ directly from our equations for a geodesic. (Note that the curves obtained by setting all but one f^i constant are geodesics, since they correspond to lines parallel to the x^i -axis.)

26. If $(V', (\cdot, \cdot)')$ and $(V'', (\cdot, \cdot)'')$ are two vector spaces with inner products, we define (\cdot, \cdot) on $V = V' \oplus V''$ by

$$(v' \oplus v'', w' \oplus w'') = (v', w')' + (v'', w'')''.$$

(a) Show that (\cdot, \cdot) is an inner product.

(b) Given Riemannian metrics on M and N , it follows that there is a natural way to put a Riemannian metric on $M \times N$. Describe the geodesics on $M \times N$ for this metric.

27. (a) Let $\gamma: [a, b] \rightarrow M$ be a geodesic, and let $p: [\alpha, \beta] \rightarrow [a, b]$ be a diffeomorphism. Show that $c = \gamma \circ p$ satisfies

$$\frac{d^2 c^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(c(t)) \frac{dc^i}{dt} \frac{dc^j}{dt} = \frac{dc^k}{dt} \frac{p''(t)}{p'(t)}.$$

(b) Conversely, if c satisfies this equation, then γ is a geodesic.

(c) If c satisfies

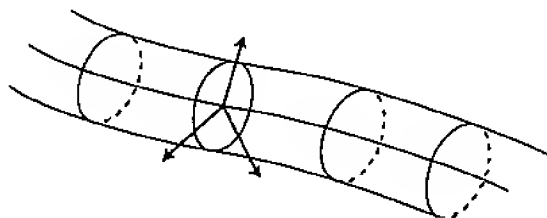
$$\frac{d^2 c^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(c(t)) \frac{dc^i}{dt} \frac{dc^j}{dt} = \frac{dc^k}{dt} \mu(t) \quad \text{for } \mu: \mathbb{R} \rightarrow \mathbb{R},$$

then c is a reparameterization of a geodesic. (The equation $p''(t) = p'(t)\mu(t)$ can be solved explicitly: $p(t) = \int^t e^{M(s)} ds$, where $M'(s) = \mu(s)$.)

28. Let c be a curve in M with $dc/dt \neq 0$ everywhere, and consider the hypersurfaces

$$\{\exp_{c(t)} v : \|v\| = \text{constant, where } v \in M_{c(t)} \text{ with } (v, dc/dt) = 0\}.$$

Show that for $v \in M_{c(t)}$ with $(v, dc/dt) = 0$, the geodesic $u \mapsto \exp_{c(t)} u \cdot v$ is perpendicular to these hypersurfaces. (Gauss' Lemma is the "special case" where c is constant.)



29. Let $\gamma: [a, b] \rightarrow M$ be a geodesic with $\gamma(a) = p$, and suppose that \exp_p is a diffeomorphism on a neighborhood $\mathcal{O} \subset M_p$ of $\{t\gamma'(0) : 0 \leq t \leq 1\}$. Show that γ is a curve of minimal length between p and $q = \gamma(b)$, among all curves in $\exp(\mathcal{O})$. (Gauss' Lemma still works on $\exp(\mathcal{O})$.)

30. If (\cdot, \cdot) is a Riemannian metric on M and $d: M \times M \rightarrow \mathbb{R}$ is the corresponding metric, then a curve $\gamma: [a, b] \rightarrow M$ with $d(\gamma(a), \gamma(b)) = L(\gamma)$ is a geodesic.

31. Schwarz's inequality for continuous functions states that

$$\left(\int_a^b fg \right)^2 \leq \left(\int_a^b f^2 \right) \left(\int_a^b g^2 \right),$$

with equality if and only if f and g are linearly dependent (over \mathbb{R}).

(a) Prove Schwarz's inequality by imitating the proof of Theorem 1(2).

(b) For any curve γ show that

$$[L_a^b(\gamma)]^2 \leq (b - a) E_a^b(\gamma),$$

with equality if and only if γ is parameterized proportionally to arclength.

(c) Let $\gamma: [a, b] \rightarrow M$ be a geodesic with $L_a^b(\gamma) = d(\gamma(a), \gamma(b))$. If $c(a) = \gamma(a)$ and $c(b) = \gamma(b)$, show that

$$E(\gamma) = \frac{L(\gamma)^2}{b - a} \leq \frac{L(c)^2}{b - a} \leq E(c).$$

Conclude that $E(\gamma) < E(c)$ unless c is also a geodesic with

$$L_a^b(c) = d(c(a), c(b)).$$

In particular, sufficiently small pieces of a geodesic minimize energy.

32. Let p be a point of a manifold M with a Riemannian metric (\cdot, \cdot) . Choose a basis v_1, \dots, v_n of M_p , so that we have a "rectangular" coordinate system χ on M_p given by $\sum_i a^i v_i \mapsto (a^1, \dots, a^n)$; let x be the coordinate system $\chi \circ \exp^{-1}$, defined in a neighborhood U of p .

(a) Show that in this coordinate system we have $\Gamma_{ij}^k(p) = 0$. *Hint:* Recall the equations for a geodesic, and note that a geodesic γ through p is just \exp composed with a straight line through 0 in M_p , so that each γ^k is linear.

(b) Let $r: U \rightarrow \mathbb{R}$ be $r(q) = d(p, q)$, so that $r \circ \gamma = \sum_k (\gamma^k)^2$. Show that

$$\frac{d^2(r \circ \gamma)^2}{dt^2} = 2 \left[\sum_k \left(\frac{d\gamma^k}{dt} \right)^2 - \sum_{i,j,k} \gamma^2 \Gamma_{ij}^k \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \right].$$

(c) Note that

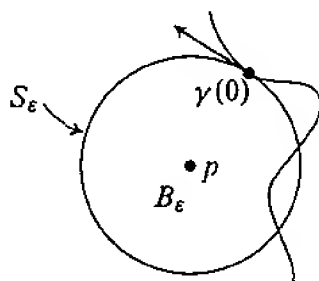
$$\sum_{i,j} \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} \leq n^2 \sum_k \left(\frac{d\gamma^k}{dt} \right)^2.$$

Using part (a), conclude that if $\|\gamma'(0)\|$ is sufficiently small, then

$$\frac{d^2(r \circ \gamma)^2}{dt^2} > 0,$$

so that $d(r \circ \gamma)^2/dt$ is strictly increasing in a neighborhood of 0.

(d) Let $B_\varepsilon = \{v \in M_p : \|v\| \leq \varepsilon\}$ and $S_\varepsilon = \{v \in M_p : \|v\| = \varepsilon\}$. Show that the following is true for all sufficiently small $\varepsilon > 0$: if γ is a geodesic such that $\gamma(0) \in \exp(S_\varepsilon)$ and such that $\gamma'(0)$ is tangent to $\exp(S_\varepsilon)$, then there is $\delta > 0$ (depending on γ) such that $\gamma(t) \notin \exp(B_\varepsilon)$ for $0 \neq t \in (-\delta, \delta)$. *Hint*: If $\gamma'(0)$ is



tangent to $\exp(S_\varepsilon)$, then $d(r \circ \gamma)/dt = 0$.

(e) Let q and q' be two points with $r(q), r(q') < \varepsilon$ and let γ be the unique geodesic of length $< 2\varepsilon$ joining them. Show that for sufficiently small ε the maximum of $r \circ \gamma$ occurs at either q or q' .

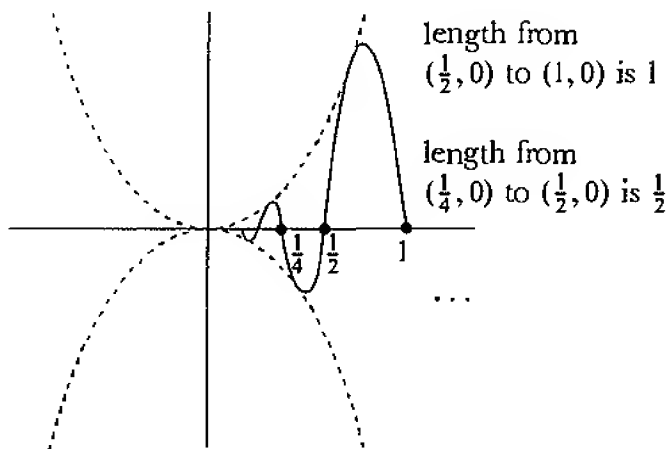
(f) A set $U \subset M$ is **geodesically convex** if every pair $q, q' \in U$ has a unique geodesic of minimum length between them, and this geodesic lies completely in U . Show that $\exp(\{v \in M_p : \|v\| < \varepsilon\})$ is geodesically convex for sufficiently small $\varepsilon > 0$.

(g) Let $f: U \rightarrow \mathbb{R}^n$ be a diffeomorphism of a neighborhood U of $0 \in \mathbb{R}^n$ into \mathbb{R}^n . Show that for sufficiently small ε , the image of the open ε -ball is convex.

33. (a) There is an everywhere differentiable curve $c(t) = (t, f(t))$ in \mathbb{R}^2 such that

$$\lim_{h \rightarrow 0} \frac{\text{length of } c|_{[0, h]}}{|c(h) - c(0)|} \neq 1.$$

Hint: Make c look something like the following picture.



(b) Consider the situation in Corollary 15, except that $c(t) = q$ if and only if $t = a$, and suppose $u'(t) > 0$ for t near 0. If c is C^1 , then $v(t)$ approaches a limit as $t \rightarrow 0$ (even though $v(0)$ is undefined). Show that if c is C^1 , then there is some $K > 0$ such that for all t near 0 we have

$$\left\| \frac{\partial \alpha}{\partial t}(u, t) \right\| \leq Ku \left\| \frac{\partial \alpha}{\partial t}(1, t) \right\| \quad 0 \leq u \leq 1.$$

Hint: In M_q we clearly have

$$\left\| \frac{d(u \cdot v(t))}{dt} \right\| = |u| \cdot \left\| \frac{dv(t)}{dt} \right\|.$$

Since \exp_q is locally a diffeomorphism there are $0 < K_1 < K_2$ such that

$$K_1 \|v\| \leq \|\exp_{q*} v\| \leq K_2 \|v\|$$

for all tangent vectors v at points near q .

(c) Conclude that

$$\lim_{h \rightarrow 0} \frac{L(c|_{[0, h]})}{d(p, c(h))} = \lim_{h \rightarrow 0} \frac{\int_0^h \sqrt{u'(t)^2 + \left\| \frac{\partial \alpha}{\partial t}(u(t), t) \right\|^2} dt}{u(h)} = 1.$$

(d) If c is C^1 , show that $L(c)$ is the least upper bound of inscribed piecewise geodesic curves.



34. (a) Using the methods of Problem 33, show that if c is the straight line joining $v, w \in M_p$, then

$$\lim_{v, w \rightarrow 0} \frac{L(\exp \circ c)}{L(c)} = 1.$$

(b) Similarly, if $\gamma_{v, w}$ is the unique geodesic joining $\exp(v)$ and $\exp(w)$, and $\gamma_{v, w} = \exp \circ c_{v, w}$, then

$$\lim_{v, w \rightarrow 0} \frac{L(\gamma_{v, w})}{L(c_{v, w})} = 1.$$

(c) Conclude that

$$\lim_{v, w \rightarrow 0} \frac{d(\exp v, \exp w)}{\|v - w\|} = 1.$$

35. Let $f: M \rightarrow N$ be an isometry. Show that f is an isometry of the metric space structures determined on M and N by their respective Riemannian metrics.

36. Let M be a manifold with Riemannian metric (\cdot, \cdot) and corresponding metric d . Let $f: M \rightarrow M$ be a map of M onto itself which preserves the metric d .

(a) If γ is a geodesic, then $f \circ \gamma$ is a geodesic.

(b) Define $f': M_p \rightarrow M_{f(p)}$ as follows: For γ a geodesic with $\gamma(0) = p$, let

$$f'(\gamma'(0)) = \left. \frac{df(\gamma(t))}{dt} \right|_{t=0}.$$

Show that $\|f'(X)\| = \|X\|$, and that $f'(cX) = cf'(X)$.

(c) Given $X, Y \in M_p$, use Problem 34 to show that

$$\begin{aligned} \frac{2(X, Y)}{\|X\| \cdot \|Y\|} &= \frac{\|X\|^2 + \|Y\|^2}{\|X\| \cdot \|Y\|} - \frac{\|tX - tY\|^2}{\|tX\| \cdot \|tY\|} \\ &= \frac{\|X\|^2 + \|Y\|^2}{\|X\| \cdot \|Y\|} - \lim_{t \rightarrow 0} \frac{[d(\exp tX, \exp tY)]^2}{\|tX\| \cdot \|tY\|}. \end{aligned}$$

Conclude that $(X, Y)_p = (f'(X), f'(Y))_{f(p)}$, and then that $f'(X + Y) = f'(X) + f'(Y)$.

(d) Part (c) shows that $f': M_p \rightarrow M_{f(p)}$ is a diffeomorphism. Use this to show that f is itself a diffeomorphism, and hence an isometry.

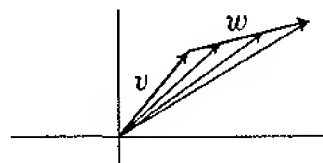
37. (a) For $v, w \in \mathbb{R}^n$ with $w \neq 0$, show that

$$\lim_{t \rightarrow 0} \frac{\|v + tw\| - \|v\|}{t} = \frac{(v, w)}{\|v\|}.$$

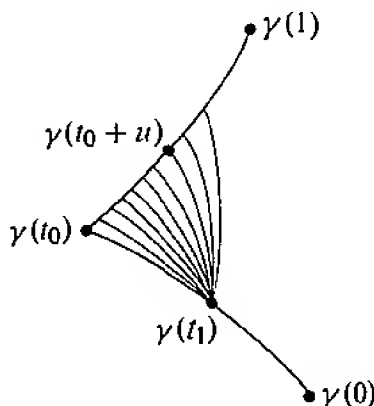
The same result then holds in any vector space with a Euclidean metric (\cdot, \cdot) .
Hint: If $\nu: \mathbb{R}^n \rightarrow \mathbb{R}$ is the norm, then the limit is $D\nu(v)(w)$. Alternately, one can use the equation $(u, v) = \|u\| \cdot \|v\| \cdot \cos \theta$ where θ is the angle between u and v .

(b) Conclude that if w is linearly independent of v , then

$$\lim_{t \rightarrow 0} \frac{\|v + tw\| - \|v\| - \|tw\|}{t} \neq 0.$$

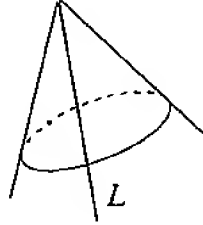


(c) Let $\gamma: [0, 1] \rightarrow M$ be a piecewise C^1 critical point for length, and suppose that $\gamma'(t_0^+) \neq \gamma'(t_0^-)$ for some $t_0 \in (0, 1)$. Choose $t_1 < t_0$ and consider the variation α for which $\tilde{\alpha}(u)$ is obtained by following γ up to t_1 , then the unique geodesic from $\gamma(t_1)$ to $\gamma(t_0 + u)$, and finally the rest of γ . Show that if t_1 is



close enough to t_0 , then $dL(\tilde{\alpha}(u))/du|_{u=0} \neq 0$, a contradiction. Thus, critical paths for length cannot have kinks.

38. Consider a cylinder $Z \subset \mathbb{R}^3$ of radius r . Find the metric d induced by the Riemannian metric it acquires as a subset of \mathbb{R}^3 .
39. Consider a cone C (without the vertex), and let L be a generating line. Unfolding $C - L$ onto \mathbb{R}^2 produces a map $f: C - L \rightarrow \mathbb{R}^2$ which is a local



isometry, but which is usually not one-one. Investigate the geodesics on a cone (the number of geodesics between two points depends on the angle of the cone, and some geodesics may come back to their initial point).

40. Let $g: S^n \rightarrow \mathbb{P}^n$ be the map $g(p) = [p] = \{p, -p\}$.
- (a) Show that there is a unique Riemannian metric $\langle \cdot, \cdot \rangle$ on \mathbb{P}^n such that $g^*\langle \cdot, \cdot \rangle$ is the usual Riemannian metric on S^n (the one that makes the inclusion of S^n into \mathbb{R}^{n+1} an isometry).
- (b) Show that every geodesic $\gamma: \mathbb{R} \rightarrow \mathbb{P}^n$ is closed (that is, there is a number a such that $\gamma(t+a) = \gamma(t)$ for all t), and that every two geodesics intersect exactly once.
- (c) Show that there are isometries of \mathbb{P}^n onto itself taking any tangent vector at one point to any tangent vector at any other point.

These results show that \mathbb{P}^n provides a model for “elliptical” non-Euclidean geometry. The sum of the angles in any triangle is $> \pi$.

41. The Poincaré upper half-plane \mathcal{H}^2 is the manifold $\{(x, y) \in \mathbb{R}^2 : y > 0\}$ with the Riemannian metric

$$\langle \cdot, \cdot \rangle = \frac{dx \otimes dx + dy \otimes dy}{y^2}.$$

- (a) Compute that

$$\Gamma_{22}^2 = \Gamma_{12}^1 = \Gamma_{21}^1 = -\frac{1}{y}, \quad \Gamma_{11}^2 = \frac{1}{y}; \quad \text{all other } \Gamma_{ij}^k = 0.$$

(b) Let C be a semi-circle in \mathcal{H}^2 with center at $(0, c)$ and radius R . Considering it as a curve $t \mapsto (t, \gamma(t))$, show that

$$\frac{d^2\gamma(t)}{dt^2} = \frac{-\gamma'(t)}{t - c} - \frac{\gamma'(t)^2}{\gamma(t)}.$$

(c) Using Problem 27, show that all the geodesics in \mathcal{H}^2 are the (suitably parameterized) semi-circles with center on the x -axis, together with the straight lines parallel to the y -axis.

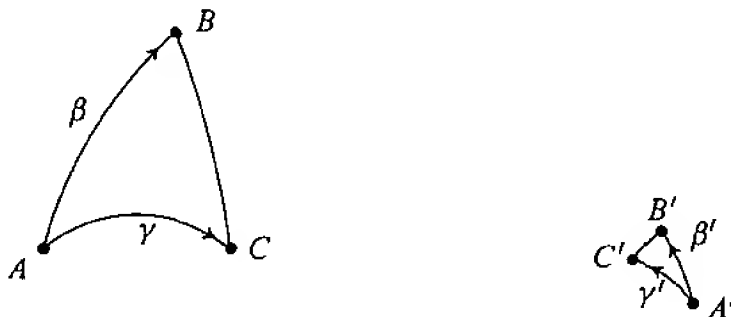
(d) Show that these geodesics have infinite length in either direction, so that the upper half-plane is complete.

(e) Show that if γ is a geodesic and $p \notin \gamma$, then there are infinitely many geodesics through p which do not intersect γ .

(f) For those who know a little about conformal mapping (compare with Problem IV.7-6). Consider the upper half-plane as a subset of the complex numbers \mathbb{C} . Show that the maps

$$f(z) = \frac{az + b}{cz + d} \quad a, b, c, d \in \mathbb{R}, \quad ad - bc > 0$$

are isometries, and that we can take any tangent vector at one point to any tangent vector at any other point by some f_* . Conclude that if length $AB =$ length $A'B'$ and length $AC =$ length $A'C'$ and the angle between the tangent vectors of β and γ at A equals the angle between the tangent vectors of β' and γ' at A' , then length $BC =$ length $B'C'$ and the angles at B and B' and at C and C' are equal ("side-angle-side"). These results show that the Poincaré



upper half-plane is a model for Lobachevskian non-Euclidean geometry. The sum of the angles in any triangle is $< \pi$.

42. Let M be a Riemannian manifold such that every two points of M can be joined by a unique geodesic of minimal length. Does it necessarily follow that the Riemannian manifold M is complete?

43. Let M be a manifold with a Riemannian metric (\cdot, \cdot) , and choose a fixed point $p \in M$. Suppose that every geodesic $\gamma: [a, b] \rightarrow M$ with initial value $\gamma(a) = p$ can be extended to all of \mathbb{R} . Show that the Riemannian manifold M is geodesically complete.

44. Let p be a point in a complete *non-compact* Riemannian manifold M . Prove that there is a geodesic $\gamma: [0, \infty) \rightarrow M$ with the initial value $\gamma(0) = p$, having the property that γ is a minimal geodesic between any two of its points.

45. Let M and N be geodesically complete Riemannian manifolds, and give $M \times N$ the Riemannian metric described in Problem 26. Show that the Riemannian manifold $M \times N$ is also complete.

46. This problem presupposes knowledge of covering spaces. Let $g: M \rightarrow N$ be a covering space, where N is a C^∞ manifold. Then there is a unique C^∞ structure on M which makes g an immersion. If (\cdot, \cdot) is a Riemannian metric on N , then $g^*(\cdot, \cdot)$ is a Riemannian metric on M , and $(M, g^*(\cdot, \cdot))$ is complete if and only if $(N, (\cdot, \cdot))$ is complete.

47. (a) If $M^n \subset N^{n+k}$ is a submanifold of N , show that the normal bundle ν is indeed a k -plane bundle.

(b) Using the notion of Whitney sum \oplus introduced in Problem 3-52, show that

$$\nu \oplus TM \simeq (TN)|_M.$$

48. (a) Show that the normal bundles ν_1, ν_2 of $M^n \subset N^{n+k}$ defined for two different Riemannian metrics are equivalent.

(b) If $\xi = \pi: E \rightarrow M$ is a smooth k -plane bundle over M^n , show that the normal bundle of $M \subset E$ is equivalent to ξ .

49. (a) Given an exact sequence of bundle maps

$$0 \longrightarrow E_1 \xrightarrow{\tilde{f}} E_2 \xrightarrow{\tilde{g}} E_3 \longrightarrow 0$$

as in Problem 3-28, where the bundles are over a smooth manifold M [or, more generally, over a paracompact space], show that $E_2 \simeq E_1 \oplus E_3$.

(b) If $\xi = \pi: E \rightarrow M$ is a smooth bundle, conclude that $TE \simeq \pi^*(\xi) \oplus \pi^*(TM)$.

50. (a) Let M be a non-orientable manifold. According to Problem 3-22 there is $S^1 \subset M$ so that $(TM)|_{S^1}$ is not orientable (the Problem deals with the case where $(TM)|_{S^1}$ is always trivial, but the same conclusions will hold if each $(TM)|_{S^1}$ is orientable; in fact, it is not hard to show that a bundle over S^1 is trivial if and only if it is orientable). Using Problem 47, show that the normal bundle ν of $S^1 \subset M$ is not orientable.
- (b) Use Problem 3-29 to conclude that there is a neighborhood of some $S^1 \subset M$ which is not orientable. (Thus, any non-orientable manifold contains a “fairly small” non-orientable open submanifold.)

CHAPTER 10

LIE GROUPS

This chapter uses, and illuminates, many of the results and concepts of the preceding chapters. It will also play an important role in later Volumes, where we are concerned with geometric problems, because in the study of these problems the groups of automorphisms of various structures play a central role, and these groups can be studied by the methods now at our disposal.

A topological group is a space G which also has a group structure (the product of $a, b \in G$ being denoted by ab) such that the maps

$$\begin{aligned}(a, b) &\mapsto ab && \text{from } G \times G \text{ to } G \\ a &\mapsto a^{-1} && \text{from } G \text{ to } G\end{aligned}$$

are continuous. It clearly suffices to assume instead that the single map $(a, b) \mapsto ab^{-1}$ is continuous. We will mainly be interested in a very special kind of topological group. A Lie group is a group G which is also a manifold with a C^∞ structure such that

$$\begin{aligned}(x, y) &\mapsto xy \\ x &\mapsto x^{-1}\end{aligned}$$

are C^∞ functions. It clearly suffices to assume that the map $(x, y) \mapsto xy^{-1}$ is C^∞ . As a matter of fact (Problem 1), it even suffices to assume that the map $(x, y) \mapsto xy$ is C^∞ .

The simplest example of a Lie group is \mathbb{R}^n , with the operation $+$. The circle S^1 is also a Lie group. One way to put a group structure on S^1 is to consider it as the quotient group \mathbb{R}/\mathbb{Z} , where $\mathbb{Z} \subset \mathbb{R}$ denotes the subgroup of integers. The functions $x \mapsto \cos 2\pi x$ and $x \mapsto \sin 2\pi x$ are C^∞ functions on \mathbb{R}/\mathbb{Z} , and at each point at least one of them is a coordinate system. Thus the map

$$\begin{array}{ccccc}(x, y) & \mapsto & x - y & \mapsto & xy^{-1} \\ \cap & & \cap & & \cap \\ \mathbb{R} \times \mathbb{R} & \longrightarrow & \mathbb{R} & \longrightarrow & S^1 = \mathbb{R}/\mathbb{Z}.\end{array}$$

which can be expressed in coordinates as one of the two maps

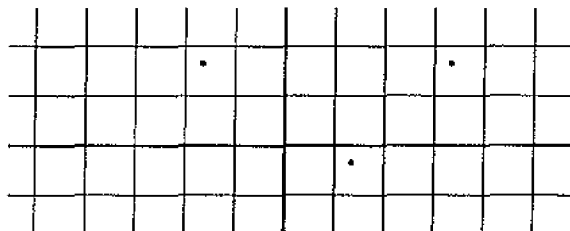
$$(x, y) \mapsto \cos 2\pi(x - y) = \cos 2\pi x \cos 2\pi y + \sin 2\pi x \sin 2\pi y$$

$$(x, y) \mapsto \sin 2\pi(x - y) = \sin 2\pi x \cos 2\pi y - \cos 2\pi x \sin 2\pi y,$$

is C^∞ : consequently the map $(x, y) \mapsto xy^{-1}$ from $S^1 \times S^1$ to S^1 is also C^∞ .

If G and H are Lie groups, then $G \times H$, with the product C^∞ structure, and the direct product group structure, is easily seen to be a Lie group. In particular, the torus $S^1 \times S^1$ is a Lie group. The torus may also be described as the quotient group

$$\mathbb{R} \times \mathbb{R} / (\mathbb{Z} \times \mathbb{Z});$$



the pairs (a, b) and (a', b') represent the same element of $S^1 \times S^1$ if and only if $a' - a \in \mathbb{Z}$ and $b' - b \in \mathbb{Z}$.

Many important Lie groups are matrix groups. The **general linear group** $GL(n, \mathbb{R})$ is the group of all non-singular real $n \times n$ matrices, considered as a subset of \mathbb{R}^{n^2} . Since the function $\det: \mathbb{R}^{n^2} \rightarrow \mathbb{R}$ is continuous (it is a polynomial map), the set $GL(n, \mathbb{R}) = \det^{-1}(\mathbb{R} - \{0\})$ is open, and hence can be given the C^∞ structure which makes it an open submanifold of \mathbb{R}^{n^2} . Multiplication of matrices is C^∞ , since the entries of AB are polynomials in the entries of A and B . Smoothness of the inverse map follows similarly from Cramer's Rule:

$$(A^{-1})_{ji} = \det A^{ij} / \det A,$$

where A^{ij} is the matrix obtained from A by deleting row i and column j .

One of the most important examples of a Lie group is the **orthogonal group** $O(n)$, consisting of all $A \in GL(n, \mathbb{R})$ with $A \cdot A^t = I$, where A^t is the transpose of A . This condition is equivalent to the condition that the rows [and columns] of A are orthonormal, which is equivalent to the condition that, with respect to the usual basis of \mathbb{R}^n , the matrix A represents a linear transformation which is an "isometry", i.e., is norm preserving, and thus inner product preserving. Problem 2-33 presents a proof that $O(n)$ is a closed submanifold of $GL(n, \mathbb{R})$, of dimension $n(n-1)/2$. To show that $O(n)$ is a Lie group we must show that the map $(x, y) \mapsto xy^{-1}$ which is C^∞ on $GL(n, \mathbb{R})$, is also C^∞ as a map from $O(n) \times O(n)$ to $O(n)$. By Proposition 2-11, it suffices to show that it is continuous; but this is true because the inclusion of $O(n) \rightarrow GL(n, \mathbb{R})$ is a homeomorphism (since $O(n)$ is a submanifold of $GL(n, \mathbb{R})$). Later in the chapter we will have another way of proving that $O(n)$ is a Lie group, and in particular, a manifold.

The argument in the previous paragraph shows, generally, that if $H \subset G$ is a subgroup of G and also a submanifold of G , then H is a Lie group. (This gives another proof that S^1 is a Lie group, for $S^1 \subset \mathbb{R}^2$ can be considered as the group

of complex numbers of norm 1. Similarly, S^3 is the Lie group of quaternions of norm 1. It is known that these are the only spheres which admit a Lie group structure.) It is possible for a subgroup H of G to be a Lie group with respect to a C^∞ structure that makes it merely an immersed submanifold. For example, if $L \subset \mathbb{R} \times \mathbb{R}$ is a subgroup consisting of all (x, cx) for c irrational, then the

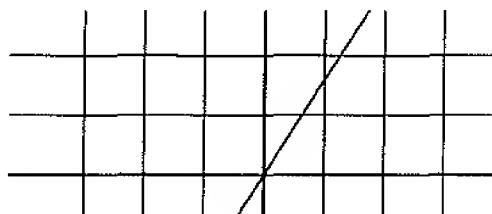


image of L in $S^1 \times S^1 = \mathbb{R} \times \mathbb{R}/(\mathbb{Z} \times \mathbb{Z})$ is a dense subgroup. We define a Lie subgroup H of G to be a subset H of G which is a subgroup of G , and also a Lie group for some C^∞ structure which makes the inclusion map $i: H \rightarrow G$ an immersion. As we have seen, a subgroup which is an (imbedded) submanifold is always a Lie subgroup. It even turns out, after some work (Problem 18), that a subgroup which is an immersed submanifold is always a Lie subgroup, but we will not need this fact.

The group $O(n)$ is disconnected; the two components consist of all $A \in O(n)$ with $\det A = +1$ and $\det A = -1$, respectively. Clearly $SO(n) = \{A \in O(n) : \det A = 1\}$, the component containing the identity I , is a subgroup. This is not accidental.

1. PROPOSITION. If G is a topological group, then the component K containing the identity $e \in G$ is a closed normal subgroup of G . If G is a Lie group, then K is an open Lie subgroup.

PROOF. If $a \in K$, then $a^{-1}K$ is connected, since $b \mapsto a^{-1}b$ is a homeomorphism of K to itself. Since $e = a^{-1}a \in a^{-1}K$, we have $a^{-1}K \subset K$. Since this is true for all $a \in K$, we have $K^{-1}K \subset K$, which proves that K is a subgroup.

For any $b \in G$, it follows similarly that bKb^{-1} is connected. Since $e \in bKb^{-1}$, we have $bKb^{-1} \subset K$, so K is normal. Moreover, K is closed since components are always closed.

If G is a Lie group, then K is also open, since G is locally connected, so K is a submanifold and a subgroup of G . Hence K is a Lie subgroup. ♦

The group $SO(2)$ is just S^1 , which we have already seen is a Lie group. As a final example of a Lie group, we mention $E(n)$, the group of all Euclidean

motions, i.e., isometries of \mathbb{R}^n . A little argument shows (Problem 5) that every element of $E(n)$ can be written uniquely as $A \cdot \tau$ where $A \in O(n)$, and τ is a translation,

$$\tau(x) = \tau_a(x) = x + a.$$

We can give $E(n)$ the C^∞ structure which makes it diffeomorphic to $O(n) \times \mathbb{R}^n$. Now $E(n)$ is not the direct product $O(n) \times \mathbb{R}^n$ as a group, since translations and orthogonal transformations do not generally commute. In fact,

$$A\tau_a A^{-1}(x) = A(A^{-1}x + a) = x + A(a) = \tau_{A(a)}(x),$$

so

$$A\tau_a A^{-1} = \tau_{A(a)}, \quad A\tau_a = \tau_{A(a)}A.$$

Consequently,

$$\begin{aligned} A\tau_a(B\tau_b)^{-1} &= A\tau_a\tau_b B^{-1} = A\tau_{a-b}B^{-1} \\ &= AB^{-1}\tau_{B(a-b)}, \end{aligned}$$

which shows that $E(n)$ is a Lie group. Clearly the component of $e \in E(n)$ is the subgroup of all $A\tau$ with $A \in SO(n)$.

For any Lie group G , if $a \in G$ we define the left and right translations, $L_a: G \rightarrow G$ and $R_a: G \rightarrow G$, by

$$\begin{aligned} L_a(b) &= ab \\ R_a(b) &= ba. \end{aligned}$$

Notice that L_a and R_a are both diffeomorphisms, with inverses $L_{a^{-1}}$ and $R_{a^{-1}}$, respectively. Consequently, the maps

$$\begin{aligned} L_{a*}: G_b &\rightarrow G_{ab} \\ R_{a*}: G_b &\rightarrow G_{ba} \end{aligned}$$

are isomorphisms. A vector field X on G is called *left invariant* if

$$L_{a*}X = X \quad \text{for all } a \in G.$$

Recall this means that

$$L_{a*}X_b = X_{ab} \quad \text{for all } a, b \in G.$$

It is easy to see that this is true if we merely have

$$L_{a*}X_e = X_a \quad \text{for all } a \in G.$$

Consequently, given $X_e \in G_e$, there is a unique left invariant vector field X on G which has the value X_e at e .

2. PROPOSITION. Every left invariant vector field X on a Lie group G is C^∞ .

PROOF. It suffices to prove that X is C^∞ in a neighborhood of e , since the diffeomorphism L_a then takes X to the C^∞ vector field $L_{a*}X$ around a (Problem 5-1). Let (x, U) be a coordinate system around e . Choose a neighborhood V of e so that $a, b \in V$ implies $ab^{-1} \in U$. Then for $a \in V$ we have

$$\begin{aligned} Xx^i(a) &= L_{a*}X_e(x^i) \\ &= X_e(x^i \circ L_a). \end{aligned}$$

Since the map $(a, b) \mapsto ab$ is C^∞ on $V \times V$ we can write

$$x^i(ab) = x^i L_a(b) = f^i(x^1(a), \dots, x^n(a), x^1(b), \dots, x^n(b))$$

for some C^∞ function f^i on $x(V) \times x(V)$. Then

$$\begin{aligned} Xx^i(a) &= X_e(x^i \circ L_a) \\ &= \sum_{j=1}^n c^j \frac{\partial(x^i \circ L_a)}{\partial x^j} \Big|_e \quad \text{where } X_e = \sum_{j=1}^n c^j \frac{\partial}{\partial x^j} \Big|_e \\ &= \sum_{j=1}^n c^j D_{n+j} f^i(x(a), x(e)). \end{aligned}$$

which shows that Xx^i is C^∞ . This implies that X is C^∞ . ♦

3. COROLLARY. A Lie group G always has a trivial tangent bundle (and is consequently orientable).

PROOF. Choose a basis X_{1e}, \dots, X_{ne} for G_e . Let X_1, \dots, X_n be the left invariant vector fields with these values at e . Then X_1, \dots, X_n are clearly everywhere linearly independent, so we can define an equivalence

$$f: TG \rightarrow G \times \mathbb{R}^n$$

by

$$f\left(\sum_{j=1}^n c^j X_j(a)\right) = (a, c^1, \dots, c^n). \quad \diamond$$

A left invariant vector field X is just one that is L_a -related to itself for all a . Consequently, Proposition 6-3 shows that $[X, Y]$ is left invariant if X and Y are. Henceforth we will use X, Y , etc., to denote elements of G_e , and \tilde{X}, \tilde{Y} , etc., to denote the left invariant vector fields with $\tilde{X}(e) = X$, $\tilde{Y}(e) = Y$, etc. We can then define an operation $[\ , \]$ on G_e by

$$[X, Y] = [\tilde{X}, \tilde{Y}](e).$$

The vector space G_e , together with this $[\ , \]$ operation, is called the **Lie algebra** of G , and will be denoted by $\mathcal{L}(G)$. (Sometimes the Lie algebra of G is defined instead to be the set of left invariant vector fields.) We will also use the more customary notation \mathfrak{g} (a German Fraktur g) for $\mathcal{L}(G)$. This notation requires some conventions for particular groups; we write

$$\begin{aligned} \mathfrak{gl}(n, \mathbb{R}) & \quad \text{for the Lie algebra of } \mathrm{GL}(n, \mathbb{R}) \\ \mathfrak{o}(n) & \quad \text{for the Lie algebra of } \mathrm{O}(n). \end{aligned}$$

In general, a **Lie algebra** is a finite dimensional vector space V , with a bilinear operation $[\ , \]$ satisfying

$$\begin{aligned} [X, X] &= 0 \\ [[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] &= 0 \quad \text{"Jacobi identity"} \end{aligned}$$

for all $X, Y, Z \in V$.

Since the $[\ , \]$ operation is assumed alternating, it is also skew-symmetric, $[X, Y] = -[Y, X]$. Consequently, we call a Lie algebra **abelian** or **commutative** if $[X, Y] = 0$ for all X, Y .

The Lie algebra of \mathbb{R}^n is isomorphic as a vector space to \mathbb{R}^n . Clearly $\mathcal{L}(\mathbb{R}^n)$ is abelian, since the vector fields $\partial/\partial x^i$ are left invariant and $[\partial/\partial x^i, \partial/\partial x^j] = 0$. The Lie algebra $\mathcal{L}(S^1)$ of S^1 is 1-dimensional, and consequently must be abelian. If V_i are Lie algebras with bracket operations $[\ , \]_i$ for $i = 1, 2$, then we can define an operation $[\ , \]$ on the direct sum $V = V_1 \oplus V_2$ ($= V_1 \times V_2$ as a set) by

$$[(X_1, X_2), (Y_1, Y_2)] = ([X_1, Y_1]_1, [X_2, Y_2]_2).$$

It is easy to check that this makes V into a Lie algebra, and that $\mathcal{L}(G \times H)$ is isomorphic to $\mathcal{L}(G) \times \mathcal{L}(H)$ with this bracket operation. Consequently, the Lie algebra $\mathcal{L}(S^1 \times \cdots \times S^1)$ is also abelian.

The structure of $\mathfrak{gl}(n, \mathbb{R})$ is more complicated. Since $\mathrm{GL}(n, \mathbb{R})$ is an open submanifold of \mathbb{R}^{n^2} , the tangent space of $\mathrm{GL}(n, \mathbb{R})$ at the identity I can be

identified with \mathbb{R}^{n^2} . If we use the standard coordinates x^{ij} on \mathbb{R}^{n^2} , then an $n \times n$ (possibly singular) matrix $M = (M_{ij})$ can be identified with

$$M_I = \sum_{i,j} M_{ij} \frac{\partial}{\partial x^{ij}} \Big|_I.$$

Let \tilde{M} be the left invariant vector field on $\mathrm{GL}(n, \mathbb{R})$ corresponding to M . We compute the function $\tilde{M}x^{kl}$ on $\mathrm{GL}(n, \mathbb{R})$ as follows. For every $A \in \mathrm{GL}(n, \mathbb{R})$,

$$\tilde{M}x^{kl}(A) = \tilde{M}_A(x^{kl}) = L_{A*}M_I(x^{kl}) = M_I(x^{kl} \circ L_A).$$

Now the function $x^{kl} \circ L_A: \mathrm{GL}(n, \mathbb{R}) \rightarrow \mathbb{R}$ is the linear function

$$(x^{kl} \circ L_A)(B) = x^{kl}(AB) = \sum_{\alpha=1}^n A_{k\alpha} B_{\alpha l},$$

with (constant) partial derivatives

$$\frac{\partial}{\partial x^{ij}}(x^{kl} \circ L_A) = \begin{cases} A_{ki} & j = l \\ 0 & j \neq l. \end{cases}$$

So

$$\begin{aligned} \tilde{M}x^{kl}(A) &= M_I(x^{kl} \circ L_A) = \sum_{i,j} M_{ij} \frac{\partial}{\partial x^{ij}}(x^{kl} \circ L_A) \\ &= \sum_{i=1}^n M_{il} A_{ki} = \sum_{\alpha=1}^n M_{\alpha l} A_{k\alpha}. \end{aligned}$$

Thus,

$$\frac{\partial}{\partial x^{ij}} \tilde{M}x^{kl} = \begin{cases} M_{jl} & k = i \\ 0 & k \neq i. \end{cases}$$

So if N is another $n \times n$ matrix, we have

$$\begin{aligned} N_I(\tilde{M}x^{kl}) &= \sum_{i,j} N_{ij} \frac{\partial}{\partial x^{ij}}(\tilde{M}x^{kl}) \\ &= \sum_{j=1}^n N_{kj} M_{jl} = (NM)_{kl}. \end{aligned}$$

From this we see that

$$[\tilde{M}, \tilde{N}]_I = \sum_{k,l} (MN - NM)_{kl} \frac{\partial}{\partial x^{kl}} \Big|_I;$$

with $\sin t$ and $-\sin t$ at (i, j) and (j, i) , 1's on the diagonal except at (i, i) and (j, j) , and 0's elsewhere. Then the set of all $A'(0)$ span the skew-symmetric matrices. Hence $O(n)_I$ must consist exactly of skew-symmetric matrices, and $O(n)$ must have dimension $n(n-1)/2$.

We do not need any new calculations to determine the bracket operation in $\mathfrak{o}(n)$. In fact, consider a Lie subgroup H of any Lie group G , and let $i: H \rightarrow G$ be the inclusion. Since $i_*: H_e \rightarrow G_e$ is an isomorphism into, we can identify H_e with a subspace of G_e . Any $X \in H_e$ can be extended to a left invariant vector field \tilde{X} on H and a left invariant vector field \tilde{X} on G . For each $a \in H \subset G$, we have left translations

$$L_a: H \rightarrow H, \quad L_a: G \rightarrow G$$

and

$$L_a \circ i = i \circ L_a.$$

So

$$i_* \tilde{X}(a) = i_* L_{a*} X = L_{a*}(i_* X) = \tilde{X}(a).$$

In other words, \tilde{X} and \tilde{X} are i -related. Consequently, if $Y \in H_e$, then $[\tilde{X}, \tilde{Y}]$ and $[\tilde{X}, \tilde{Y}]$ are i -related, which means that

$$[\tilde{X}, \tilde{Y}](e) = i_*([\tilde{X}, \tilde{Y}](e)).$$

Thus, $H_e \subset G_e = \mathfrak{g}$ is a subalgebra of \mathfrak{g} , that is, H_e is a subspace of \mathfrak{g} which is closed under the $[\ , \]$ operation; moreover, H_e with this induced $[\ , \]$ operation is just $\mathfrak{h} = \mathcal{L}(H)$.

This correspondence between Lie subgroups of G and subalgebras of \mathfrak{g} turns out to work in the other direction also.

4. THEOREM. Let G be a Lie group, and \mathfrak{h} a subalgebra of \mathfrak{g} . Then there is a unique connected Lie subgroup H of G whose Lie algebra is \mathfrak{h} .

PROOF. For $a \in G$, let Δ_a be the subspace of G_a consisting of all $\tilde{X}(a)$ for $X \in \mathfrak{h}$. The fact that \mathfrak{h} is a subalgebra of \mathfrak{g} implies that Δ is an integrable distribution. Let H be the maximal integral manifold of Δ containing e . If $b \in G$, then clearly $L_{b*}(\Delta_a) = \Delta_{ba}$, so L_{b*} leaves the distribution Δ invariant. It follows immediately that L_b permutes the various maximal integral manifolds of Δ among themselves. In particular, if $b \in H$, then $L_{b^{-1}}$ takes H to the maximal integral manifold containing $L_{b^{-1}}(b) = e$, so $L_{b^{-1}}(H) = H$. This implies that H is a subgroup of G . To prove that it is a Lie subgroup we just need to show that $(a, b) \mapsto ab^{-1}$ is C^∞ . Now this map is clearly C^∞ as a map into G . Using Theorem 6-7, it follows that it is C^∞ as a map into H .

The proof of uniqueness is left to the reader. ♦

There is a very difficult theorem of Ado which states that every Lie algebra is isomorphic to a subalgebra of $\text{GL}(N, \mathbb{R})$ for some N . It then follows from Theorem 4 that *every Lie algebra is isomorphic to the Lie algebra of some Lie group*. Later on we will be able to obtain a “local” version of this result. We will soon see to what extent the Lie algebra of G determines G .

We continue the study of Lie groups along the same route used in the study of groups. Having considered subgroups of Lie groups (and subalgebras of their Lie algebras), we next consider, more generally, homomorphisms between Lie groups. If $\phi: G \rightarrow H$ is a C^∞ homomorphism, then $\phi_{*e}: G_e \rightarrow H_e$. For any $a \in G$ we clearly have

$$\phi \circ L_a = L_{\phi(a)} \circ \phi,$$

so if $X \in G_e$, and $\tilde{X} = \widetilde{\phi_{*e}X}$ is the left invariant vector field on H with value $\phi_{*e}X$ at e , then

$$\begin{aligned}\phi_{*a}\tilde{X}(a) &= \phi_{*a}L_{a*}X = L_{\phi(a)*}\phi_{*e}X \\ &= \tilde{X}(\phi(a)).\end{aligned}$$

Thus \tilde{X} and \tilde{Y} are ϕ -related. Consequently, the map $\phi_{*e}: \mathfrak{g} \rightarrow \mathfrak{h}$ is a Lie algebra homomorphism, that is,

$$\begin{aligned}\phi_{*e}(aX + bY) &= a\phi_{*e}X + b\phi_{*e}Y \\ \phi_{*e}[X, Y] &= [\phi_{*e}X, \phi_{*e}Y].\end{aligned}$$

Usually, we will denote ϕ_{*e} simply by $\phi_*: \mathfrak{g} \rightarrow \mathfrak{h}$.

For example, suppose that $G = H = \mathbb{R}$. There are an enormous number of homomorphisms $\phi: \mathbb{R} \rightarrow \mathbb{R}$, because \mathbb{R} is a vector space of uncountable dimension over \mathbb{Q} , and every linear transformation is a group homomorphism. But if ϕ is C^∞ , then the condition

$$\phi(s + t) = \phi(s) + \phi(t)$$

implies that

$$\frac{d\phi(t+s)}{ds} = \frac{d\phi(s)}{ds};$$

evaluating at $s = 0$ gives

$$\phi'(t) = \phi'(0),$$

which means that $\phi(t) = ct$ for some $c (= \phi'(0))$. It is not hard to see that even a continuous ϕ must be of this form (one first shows that ϕ is of this form on the

rational numbers). We can identify $\mathcal{L}(\mathbb{R})$ with \mathbb{R} . Clearly the map $\phi_*: \mathbb{R} \rightarrow \mathbb{R}$ is just multiplication by c .

Now suppose that $G = \mathbb{R}$, but $H = S^1 = \mathbb{R}/\mathbb{Z}$. A neighborhood of the identity $e \in S^1$ can be identified with a neighborhood of $0 \in \mathbb{R}$, giving rise to an identification of $\mathcal{L}(S^1)$ with \mathbb{R} . The continuous homomorphisms $\phi: \mathbb{R} \rightarrow S^1$ are clearly of the form

$$\mathbb{R} \xrightarrow{\times c} \mathbb{R} \longrightarrow \mathbb{R}/\mathbb{Z};$$

once again, $\phi_*: \mathbb{R} \rightarrow \mathbb{R}$ is multiplication by c .

Notice that the only continuous homomorphism $\phi: S^1 \rightarrow \mathbb{R}$ is the 0 map (since $\{0\}$ is the only compact subgroup of \mathbb{R}). Consequently, a Lie algebra homomorphism $\mathfrak{g} \rightarrow \mathfrak{h}$ may not come from any C^∞ homomorphism $\phi: G \rightarrow H$. However, we do have a local result.

5. THEOREM. Let G and H be Lie groups, and $\Phi: \mathfrak{g} \rightarrow \mathfrak{h}$ a Lie algebra homomorphism. Then there is a neighborhood U of $e \in G$ and a C^∞ map $\phi: U \rightarrow H$ such that

$$\phi(ab) = \phi(a)\phi(b) \quad \text{when } a, b, ab \in U,$$

and such that for every $X \in \mathfrak{g}$ we have

$$\phi_{*e}X = \Phi(X).$$

Moreover, if there are two C^∞ homomorphisms $\phi, \psi: G \rightarrow H$ with $\phi_{*e} = \psi_{*e} = \Phi$, and G is connected, then $\phi = \psi$.

PROOF. Let \mathfrak{k} (German Fraktur k) be the subset $\mathfrak{k} \subset \mathfrak{g} \times \mathfrak{h}$ of all $(X, \Phi(X))$, for $X \in \mathfrak{g}$. Since Φ is a homomorphism, \mathfrak{k} is a subalgebra of $\mathfrak{g} \times \mathfrak{h} = \mathcal{L}(G \times H)$. By Theorem 4, there is a unique connected Lie subgroup K of $G \times H$ whose Lie algebra is \mathfrak{k} . If $\pi_1: G \times H \rightarrow G$ is projection on the first factor, and $\omega = \pi_1|_K$, then $\omega: K \rightarrow G$ is a C^∞ homomorphism. For $X \in \mathfrak{g}$ we have

$$\omega_*(X, \Phi(X)) = X,$$

so $\omega_*: K_{(e,e)} \rightarrow G_e$ is an isomorphism. Consequently, there is an open neighborhood V of $(e, e) \in K$ such that ω takes V diffeomorphically onto an open neighborhood U of $e \in G$. If $\pi_2: G \times H \rightarrow H$ is projection on the second factor, we can define

$$\phi = \pi_2 \circ \omega^{-1} \quad \text{on } U.$$

The first condition on ϕ is obvious. As for the second, if $X \in \mathfrak{g}$, then

$$\omega_*(X, \Phi(X)) = X,$$

so

$$\phi_*X = \pi_{2*}(X, \Phi(X)) = \Phi(X).$$

Given $\phi, \psi: G \rightarrow H$, define the one-one map $\theta: G \rightarrow G \times H$ by

$$\theta(a) = (a, \psi(a)).$$

The image G' of θ is a Lie subgroup of $G \times H$ and for $X \in \mathfrak{g}$ we clearly have

$$\theta_*X = (X, \Phi(X)),$$

so $\mathcal{L}(G') = \mathfrak{f}$. Thus $G' = K$, which implies that $\psi(a) = \phi(a)$ for all $a \in G$. ♦

6. COROLLARY. If two Lie groups G and H have isomorphic Lie algebras, then they are locally isomorphic.

PROOF. Given an isomorphism $\Phi: \mathfrak{g} \rightarrow \mathfrak{h}$, let ϕ be the map given by Theorem 5. Since $\phi_{*e} = \Phi$ is an isomorphism, ϕ is a diffeomorphism in a neighborhood of $e \in G$. ♦

Remark: For those who know about simply-connected spaces it is fairly easy (Problem 8) to conclude that two simply-connected Lie groups with isomorphic Lie algebras are actually isomorphic, and that all connected Lie groups with a given Lie algebra are covered by the same simply-connected Lie group.

7. COROLLARY. A connected Lie group G with an abelian Lie algebra is itself abelian.

PROOF. By Corollary 6, G is locally isomorphic to \mathbb{R}^n , so $ab = ba$ for a, b in a neighborhood of e . It follows that G is abelian, since (Problem 4) any neighborhood of e generates G . ♦

8. COROLLARY. For every $X \in G_e$, there is a unique C^∞ homomorphism $\phi: \mathbb{R} \rightarrow G$ such that

$$\left. \frac{d\phi}{dt} \right|_{t=0} = X.$$

FIRST PROOF. Define $\Phi: \mathbb{R} \rightarrow \mathcal{L}(G)$ by

$$\Phi(\alpha) = \alpha X.$$

Clearly Φ is a Lie algebra homomorphism. By Theorem 5, on some neighborhood $(-\varepsilon, \varepsilon)$ of $0 \in \mathbb{R}$ there is a map $\phi: (-\varepsilon, \varepsilon) \rightarrow G$ with

$$\phi(s+t) = \phi(s)\phi(t) \quad |s|, |t|, |s+t| < \varepsilon$$

and

$$\left. \frac{d\phi}{dt} \right|_{t=0} = \phi_* \left(\left. \frac{d}{dt} \right|_{t=0} \right) = X.$$

To extend ϕ to \mathbb{R} we write every t with $|t| \geq \varepsilon$ uniquely as

$$t = k(\varepsilon/2) + r \quad k \text{ an integer, } |r| < \varepsilon/2$$

and define

$$\phi(t) = \begin{cases} \phi(\varepsilon/2) \cdots \phi(\varepsilon/2) \cdot \phi(r) & [\phi(\varepsilon/2) \text{ appears } k \text{ times}] \quad k \geq 0 \\ \phi(-\varepsilon/2) \cdots \phi(-\varepsilon/2) \cdot \phi(r) & [\phi(-\varepsilon/2) \text{ appears } -k \text{ times}] \quad k < 0. \end{cases}$$

Uniqueness also follows from Theorem 5.

SECOND (DIRECT) PROOF. If $f: G \rightarrow \mathbb{R}$ is C^∞ , and $\phi: \mathbb{R} \rightarrow G$ is a C^∞ homomorphism, then

$$\begin{aligned} \frac{d\phi}{dt}(f) &= \lim_{h \rightarrow 0} \frac{f(\phi(t+h)) - f(\phi(t))}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(\phi(t)\phi(h)) - f(\phi(t))}{h} \\ &= \left. \frac{d}{du} \right|_{u=0} f \circ L_{\phi(t)} \circ \phi \\ &= L_{\phi(t)*} \left. \frac{d\phi}{du} \right|_{u=0} (f) \\ &= L_{\phi(t)*} X(f) = \tilde{X}(\phi(t))(f). \end{aligned}$$

Thus ϕ must be an integral curve of \tilde{X} , which proves uniqueness. Conversely, if $\phi: \mathbb{R} \rightarrow G$ is an integral curve of \tilde{X} , then

$$t \mapsto \phi(s) \cdot \phi(t)$$

is an integral curve of \tilde{X} which passes through $\phi(s)$ at time $t = 0$. The same is clearly true for

$$t \mapsto \phi(s+t),$$

so ϕ is a homomorphism. We know that integral curves of \tilde{X} exist locally; they can be extended to all of \mathbb{R} using the method of the first proof. ♦

A homomorphism $\phi: \mathbb{R} \rightarrow G$ is called a **1-parameter subgroup** of G . We thus see that there is a unique 1-parameter subgroup ϕ of G with given tangent vector $d\phi/dt(0) \in G_e$. We have already examined the 1-parameter subgroups of \mathbb{R} . More interesting things happen when we take G to be $\mathbb{R} - \{0\}$, with multiplication as the group operation. Then all C^∞ homomorphisms $\phi: \mathbb{R} \rightarrow \mathbb{R} - \{0\}$, with

$$\phi(s+t) = \phi(s)\phi(t),$$

must satisfy

$$\phi'(t) = \phi'(0)\phi(t)$$

$$\phi(0) = 1.$$

The solutions of this equation are

$$\phi(t) = e^{\phi'(0)t}.$$

Notice that $\mathbb{R} - \{0\}$ is just $GL(1, \mathbb{R})$. All C^∞ homomorphisms $\phi: \mathbb{R} \rightarrow GL(n, \mathbb{R})$ must satisfy the analogous differential equation

$$\begin{aligned} (*) \quad \phi'(t) &= \phi'(0) \cdot \phi(t), \\ \phi(0) &= I, \end{aligned}$$

where \cdot now denotes matrix multiplication. The solutions of these equations can be written formally in the same way

$$(**) \quad \phi(t) = \exp(t\phi'(0)),$$

where exponentiation of matrices is defined by

$$\exp(A) = I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots.$$

This follows from the facts in Problem 5-6, some of which will be briefly recapitulated here.

If $A = (a_{ij})$ and $|A| = \max |a_{ij}|$, then clearly

$$\begin{aligned} |A+B| &\leq |A| + |B| \\ |AB| &\leq n|A| \cdot |B|; \end{aligned}$$

hence $|A|^k \leq n^{k-1}|A|^k \leq n^k|A|^k$. Consequently,

$$\left| \frac{A^N}{N!} + \cdots + \frac{A^{N+K}}{(N+K)!} \right| \leq \frac{(n|A|)^N}{N!} + \cdots + \frac{(n|A|)^{N+K}}{(N+K)!} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

so the series for $\exp(A)$ converges (the $(i, j)^{\text{th}}$ entry of the partial sums converge), and convergence is absolute and uniform in any bounded set. Moreover (see Problem 5-6), if $AB = BA$, then

$$\exp(A + B) = (\exp A)(\exp B).$$

Hence, if $\phi(t)$ is defined by (**), then

$$\begin{aligned}\phi'(t) &= \lim_{h \rightarrow 0} \frac{\exp(t\phi'(0) + h\phi'(0)) - \exp(t\phi'(0))}{h} \\ &= \lim_{h \rightarrow 0} \frac{[\exp(h\phi'(0)) - I]}{h} \exp(t\phi'(0)) \\ &= \lim_{h \rightarrow 0} \frac{\frac{h\phi'(0)}{1!} + \frac{h^2\phi'(0)^2}{2!} + \dots}{h} \exp(t\phi'(0)) \\ &= \phi'(0)\phi(t),\end{aligned}$$

so ϕ does satisfy (*).

For any Lie group G , we now define the “exponential map”

$$\exp: \mathfrak{g} \rightarrow G$$

as follows. Given $X \in \mathfrak{g}$, let $\phi: \mathbb{R} \rightarrow G$ be the unique C^∞ homomorphism with $d\phi/dt(0) = X$. Then

$$\exp(X) = \phi(1).$$

We clearly have

$$\begin{aligned}\exp(t_1 + t_2)X &= (\exp t_1 X)(\exp t_2 X) \\ \exp(-tX) &= (\exp tX)^{-1}.\end{aligned}$$

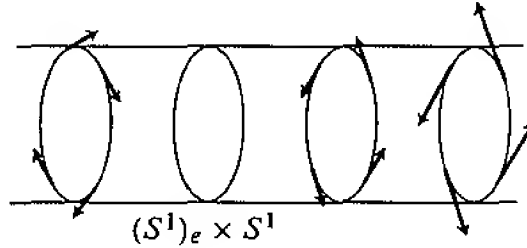
9. PROPOSITION. The map $\exp: G_e \rightarrow G$ is C^∞ (note that $G_e \approx \mathbb{R}^n$ has a natural C^∞ structure), and 0 is a regular point, so that \exp takes a neighborhood of $0 \in G_e$ diffeomorphically onto a neighborhood of $e \in G$. If $\psi: G \rightarrow H$ is any C^∞ homomorphism, then

$$\exp \circ \psi_* = \psi \circ \exp.$$

$$\begin{array}{ccc} G_e & \xrightarrow{\psi_*} & H_e \\ \exp \downarrow & & \downarrow \exp \\ G & \xrightarrow{\psi} & H \end{array}$$

PROOF. The tangent space $(G_e \times G)_{(X,a)}$ of the C^∞ manifold $G_e \times G$ at the point (X,a) can be identified with $G_e \oplus G_a$. We define a vector field Y on $G_e \times G$ by

$$Y_{(X,a)} = 0 \oplus \tilde{X}(a).$$



Then Y has a flow $\alpha: \mathbb{R} \times (G_e \times G) \rightarrow G_e \times G$, which we know is C^∞ . Since

$$\exp X = \text{projection on } G \text{ of } \alpha(1, 0 \oplus X),$$

it follows that \exp is C^∞ .

If we identify a vector $v \in (G_e)_0$ with G_e , then the curve $c(t) = tv$ in G_e has tangent vector v at 0. So

$$\begin{aligned} \exp_{*0}(v) &= \left. \frac{d \exp(c(t))}{dt} \right|_{t=0} = \left. \frac{d}{dt} \right|_{t=0} \exp(tv) \\ &= v. \end{aligned}$$

So \exp_{*0} is the identity, and hence one-one. Therefore \exp is a diffeomorphism in a neighborhood of 0.

Given $\psi: G \rightarrow H$, and $X \in G_e$, let $\phi: \mathbb{R} \rightarrow G$ be a homomorphism with

$$\left. \frac{d\phi}{dt} \right|_{t=0} = X.$$

Then $\psi \circ \phi: \mathbb{R} \rightarrow H$ is a homomorphism with

$$\left. \frac{d(\psi \circ \phi)}{dt} \right|_{t=0} = \psi_* X.$$

Consequently,

$$\exp(\psi_* X) = \psi \circ \phi(1) = \psi(\exp X). \quad \spadesuit$$

10. COROLLARY. Every one-one C^∞ homomorphism $\phi: G \rightarrow H$ is an immersion (so $\phi(G)$ is a Lie subgroup of H).

PROOF. If $\phi_{*p}(\tilde{X}(p)) = 0$ for some non-zero $X \in \mathfrak{g}$, then also $\phi_{*e}(X) = 0$. But then

$$e = \exp \phi_{*e}(tX) = \phi(\exp(tX)),$$

contradicting the fact that ϕ is one-one. \spadesuit

11. COROLLARY. Every continuous homomorphism $\phi: \mathbb{R} \rightarrow G$ is C^∞ .

PROOF. Let U be a star-shaped open neighborhood of $0 \in G_e$ on which \exp is one-one. For any $a \in \exp(\frac{1}{2}U)$, if $a = \exp(X/2)$ for $X \in U$, then

$$a = \exp(X/2) = [\exp(X/4)]^2, \quad \exp X/4 \in \exp(\tfrac{1}{2}U).$$

So a has a square root in $\exp(\frac{1}{2}U)$. Moreover, if $a = b^2$ for $b \in \exp(\frac{1}{2}U)$, then $b = \exp(Y/2)$ for $Y \in U$, so

$$\exp(X/2) = a = b^2 = [\exp(Y/2)]^2 = \exp Y.$$

Since $X/2, Y \in U$ it follows that $X/2 = Y$, so $X/4 = Y/2$. This shows that every $a \in \exp(\frac{1}{2}U)$ has a unique square root in the set $\exp(\frac{1}{2}U)$.

Now choose $\varepsilon > 0$ so that $\phi(t) \in \exp(\frac{1}{2}U)$ for $|t| \leq \varepsilon$. Let $\phi(\varepsilon) = \exp X$, $X \in \exp(\frac{1}{2}U)$. Since

$$[\phi(\varepsilon/2)]^2 = \phi(\varepsilon) = [\exp X/2]^2,$$

it follows from the above that $\phi(\varepsilon/2) = \exp(X/2)$. By induction we have

$$\phi(\varepsilon/2^n) = \exp(X/2^n).$$

Hence

$$\phi(m/2^n \cdot \varepsilon) = \phi(\varepsilon/2^n)^m = [\exp(X/2^n)]^m = \exp(m/2^n \cdot X).$$

By continuity,

$$\phi(s\varepsilon) = \exp sX \quad \text{for all } s \in [-1, 1]. \quad \spadesuit$$

12. COROLLARY. Every continuous homomorphism $\phi: G \rightarrow H$ is C^∞ .

PROOF. Choose a basis X_1, \dots, X_n for G_e . The map $t \mapsto \phi(\exp tX_i)$ is a continuous homomorphism of \mathbb{R} to H , so there is $Y_i \in H_e$ such that

$$\phi(\exp tX_i) = \exp tY_i.$$

Thus,

$$(*) \quad \phi((\exp t_1 X_1) \cdots (\exp t_n X_n)) = (\exp t_1 Y_1) \cdots (\exp t_n Y_n).$$

Now the map $\psi: \mathbb{R}^n \rightarrow G$ given by

$$\psi(t_1, \dots, t_n) = (\exp t_1 X_1) \cdots (\exp t_n X_n)$$

is C^∞ and clearly

$$\psi_* \left(\frac{\partial}{\partial x^i} \Big|_0 \right) = X_i,$$

so ψ is a diffeomorphism of a neighborhood U of $0 \in \mathbb{R}^n$ onto a neighborhood V of $e \in G$. Then on V ,

$$\phi = (\phi \circ \psi) \circ \psi^{-1},$$

and $(*)$ shows that $\phi \circ \psi$ is C^∞ . So ϕ is C^∞ at e , and thus everywhere. \spadesuit

13. COROLLARY. If G and G' are Lie groups which are isomorphic as topological groups, then they are isomorphic as Lie groups, that is, there is a diffeomorphism between them which is also a group isomorphism.

PROOF. Apply Corollary 11 to the continuous isomorphism and its inverse. ♦

The properties of the particular exponential map

$$\exp: \mathbb{R}^{n^2} (= \mathfrak{gl}(n, \mathbb{R})) \rightarrow \mathrm{GL}(n, \mathbb{R})$$

may now be used to show that $\mathrm{O}(n)$ is a Lie group. It is easy to see that

$$\exp(M^t) = (\exp M)^t.$$

Moreover, since $\exp(M + N) = (\exp M)(\exp N)$ when $MN = NM$, we have

$$(\exp M)(\exp -M) = I.$$

So if M is skew-symmetric, $M = -M^t$, then

$$(\exp M)(\exp M)^t = I,$$

i.e., $\exp M \in \mathrm{O}(n)$. Conversely, any $A \in \mathrm{O}(n)$ sufficiently close to I can be written $A = \exp M$ for some M . Let $A^t = \exp N$. Then $I = A \cdot A^t = (\exp M)(\exp N)$, so $\exp N = (\exp M)^{-1} = \exp(-M)$. For sufficiently small M and N this implies that $N = -M$. So $\exp M^t = A^t = \exp(-M)$; hence $M^t = -M$. It follows that a neighborhood of I in $\mathrm{O}(n)$ is an $n(n-1)/2$ dimensional submanifold of $\mathrm{GL}(n, \mathbb{R})$. Since $\mathrm{O}(n)$ is a subgroup, $\mathrm{O}(n)$ is itself a submanifold of $\mathrm{GL}(n, \mathbb{R})$.

Just as in $\mathrm{GL}(n, \mathbb{R})$, the equation $\exp(X + Y) = \exp X \exp Y$ holds whenever $[X, Y] = 0$ (Problem 13). In general, $[X, Y]$ measures, up to first order, the extent to which this equation fails to hold. In the following Theorem, and in its proof, to indicate that a function $c: \mathbb{R} \rightarrow G_e$ has the property that $c(t)/t^3$ is bounded for small t , we will denote it by $O(t^3)$. Thus $O(t^3)$ will denote different functions at different times.

14. THEOREM. If G is a Lie group and $X, Y \in G_e$, then

- (1) $\exp tX \exp tY = \exp \left\{ t(X + Y) + \frac{t^2}{2}[X, Y] + O(t^3) \right\}$
- (2) $\exp(-tX) \exp(-tY) \exp tX \exp tY = \exp \{ t^2[X, Y] + O(t^3) \}$
- (3) $\exp tX \exp tY \exp(-tX) = \exp \{ tY + t^2[X, Y] + O(t^3) \}.$

PROOF. We have

$$(i) \quad \tilde{X}f(a) = \tilde{X}_a(f) = L_{a*}X(f) = X(f \circ L_a) = \left. \frac{d}{du} \right|_{u=0} f(a \cdot \exp uX).$$

Similarly,

$$(ii) \quad \tilde{Y}f(a) = \left. \frac{d}{du} \right|_{u=0} f(a \cdot \exp uY).$$

For fixed s , let

$$\phi(t) = f(\exp sX \exp tY).$$

Then

$$(iii) \quad \begin{aligned} \phi'(t) &= \frac{d}{dt} f(\exp sX \exp tY) = \left. \frac{d}{du} \right|_{u=0} f(\exp sX \exp tY \exp uY) \\ &= (\tilde{Y}f)(\exp sX \exp tY) \quad \text{by (ii).} \end{aligned}$$

Applying (iii) to $\tilde{Y}f$ instead of f gives

$$(iv) \quad \phi''(t) = [\tilde{Y}(\tilde{Y}f)](\exp sX \exp tY).$$

Now Taylor's Theorem says that

$$\phi(t) = \phi(0) + \phi'(0)t + \frac{\phi''(0)}{2!}t^2 + O(t^3).$$

Suppose that $f(e) = 0$. Then we have

$$(v) \quad \begin{aligned} f(\exp sX \exp tY) &= f(\exp sX) + t(\tilde{Y}f)(\exp sX) \\ &\quad + \frac{t^2}{2}[\tilde{Y}(\tilde{Y}f)](\exp sX) + O(t^3). \end{aligned}$$

Similarly, for any F .

$$\begin{aligned} \frac{d}{ds} F(\exp sX) &= (\tilde{X}F)(\exp sX) \\ \frac{d^2}{ds^2} F(\exp sX) &= [\tilde{X}(\tilde{X}F)](\exp sX) \\ F(\exp sX) &= F(e) + s(\tilde{X}F)(e) + \frac{s^2}{2}[\tilde{X}(\tilde{X}F)](e) + O(s^3). \end{aligned}$$

Substituting in (v) for $F = f$, $F = \tilde{Y}f$, and $F = \tilde{Y}(\tilde{Y}f)$ gives

$$\begin{aligned}
 \text{(vi)} \quad f(\exp sX \exp tY) &= s(\tilde{X}f)(e) + t(\tilde{Y}f)(e) \\
 &\quad + \frac{s^2}{2}[\tilde{X}(\tilde{X}f)](e) + \frac{t^2}{2}[\tilde{Y}(\tilde{Y}f)](e) + st\tilde{X}(\tilde{Y}f)(e) \\
 &\quad + O(s^3) + O(t^3) + O(s^2t) + O(st^2).
 \end{aligned}$$

In particular,

$$\begin{aligned}
 \text{(vii)} \quad f(\exp tX \exp tY) &= t[(\tilde{X} + \tilde{Y})f](e) \\
 &\quad + t^2 \left[\left(\frac{\tilde{X}\tilde{X}}{2} + \tilde{X}\tilde{Y} + \frac{\tilde{Y}\tilde{Y}}{2} \right) f \right](e) + O(t^3).
 \end{aligned}$$

Now for small t we can write

$$\exp tX \exp tY = \exp Z(t)$$

for some C^∞ function Z with values in G_e . Applying Taylor's formula to Z gives

$$Z(t) = tZ_1 + t^2Z_2 + O(t^3),$$

for some $Z_1, Z_2 \in G_e$. If $f(e) = 0$, then clearly $f(A(t) + O(t^3)) = f(A(t)) + O(t^3)$, so by (vi) we have

$$\begin{aligned}
 \text{(viii)} \quad f(\exp Z(t)) &= f(\exp(tZ_1 + t^2Z_2)) + O(t^3) \\
 &= t(\tilde{Z}_1f)(e) + t^2(\tilde{Z}_2f)(e) \\
 &\quad + \frac{t^2}{2}[\tilde{Z}_1(\tilde{Z}_1f)](e) + O(t^3).
 \end{aligned}$$

Since we can take the f 's to be coordinate functions, comparison of (vii) and (viii) gives

$$\begin{aligned}
 \tilde{X} + \tilde{Y} &= \tilde{Z}_1 \\
 \frac{\tilde{Z}_1\tilde{Z}_1}{2} + \tilde{Z}_2 &= \frac{\tilde{X}\tilde{X}}{2} + \tilde{X}\tilde{Y} + \frac{\tilde{Y}\tilde{Y}}{2}.
 \end{aligned}$$

which gives

$$Z_1 = X + Y, \quad Z_2 = \frac{1}{2}[X, Y],$$

thus proving (1).

Equation (2) follows immediately from (1).

To prove (3), again choose f with $f(e) = 0$. Then similar calculations give

$$\begin{aligned}
 \text{(ix)} \quad & f(\exp tX \exp tY \exp(-tX)) \\
 &= t[(\tilde{X} + \tilde{Y} - \tilde{X})f](e) + t^2 \left[\left(\frac{\tilde{X}\tilde{X}}{2} + \frac{\tilde{Y}\tilde{Y}}{2} + \frac{\tilde{X}\tilde{X}}{2} + \tilde{X}\tilde{Y} - \tilde{X}\tilde{X} - \tilde{Y}\tilde{X} \right) \right](e) \\
 &\quad + O(t^3).
 \end{aligned}$$

If we write

$$\exp tX \exp tY \exp(-tX) = \exp(tS_1 + t^2S_2 + O(t^3)),$$

then we also have

$$\begin{aligned}
 \text{(x)} \quad & f(\exp tX \exp tY \exp(-tX)) = f(\exp(tS_1 + t^2S_2)) + O(t^3) \\
 &= t(\tilde{S}_1 f)(e) + t^2(\tilde{S}_2 f)(e) \\
 &\quad + \frac{t^2}{2}[\tilde{S}_1(\tilde{S}_1 f)](e) + O(t^3).
 \end{aligned}$$

Comparing (ix) and (x) gives the desired result. ♦

Notice that formula (2) is a special case of Theorem 5-16 (compare also with Problems 5-16 and 5-18).

The work involved in proving Theorem 14 is justified by its role in the following beautiful theorem.

15. THEOREM. If G is a Lie group and $H \subset G$ is a closed subset which is also a subgroup (algebraically), then H is a Lie subgroup of G . More precisely, there is a C^∞ structure on H , with the relative topology, that makes it a Lie subgroup of G .

PROOF. We attempt to reconstruct the Lie algebra of H as follows. Let $\mathfrak{h} \subset G_e$ be the set of all $X \in G_e$ such that $\exp tX \in H$ for all t .

Assertion 1. Let $X_i \in G_e$ with $X_i \rightarrow X$ and let $t_i \rightarrow 0$ with each $t_i \neq 0$. Suppose $\exp t_i X_i \in H$ for all i . Then $X \in \mathfrak{h}$.

Proof. We can assume $t_i > 0$, since $\exp(-t_i X_i) = (\exp t_i X_i)^{-1} \in H$. For $t > 0$, let

$$k_i(t) = \text{largest integer} \leq \frac{t}{t_i}.$$

Then

$$\frac{t}{t_i} - 1 < k_i(t) \leq \frac{t}{t_i},$$

so

$$t_i k_i(t) \rightarrow t.$$

Now

$$\begin{aligned} \exp(k_i(t)t_i X_i) &= [\exp(t_i X_i)]^{k_i(t)} \in H, \\ k_i(t)t_i X_i &\rightarrow tX. \end{aligned}$$

Thus $\exp tX \in H$, since H is closed and \exp is continuous. We clearly also have $\exp tX \in H$ for $t < 0$, so $X \in \mathfrak{h}$. Q.E.D.

We now claim that $\mathfrak{h} \subset G_e$ is a vector subspace. Clearly $X \in \mathfrak{h}$ implies $sX \in \mathfrak{h}$ for all $s \in \mathbb{R}$. If $X, Y \in \mathfrak{h}$, we can write by (1) of Theorem 14

$$\exp tX \exp tY = \exp\{t(X + Y) + tZ(t)\}$$

where $Z(t) \rightarrow 0$ as $t \rightarrow 0$. Choose positive $t_i \rightarrow 0$ and let $X_i = X + Y + Z(t_i)$. Then *Assertion 1* implies that $X + Y \in \mathfrak{h}$. Alternatively, we can write, for fixed t ,

$$\left(\exp \frac{t}{n} X \exp \frac{t}{n} Y\right)^n = \exp \left\{ t(X + Y) + \frac{t^2}{2n} [X, Y] + O(1/n^2) \right\};$$

taking limits as $n \rightarrow \infty$ gives $\exp t(X + Y) \in H$.

[Similarly, using (2) of Theorem 14 we see that $[X, Y] \in \mathfrak{h}$, so that \mathfrak{h} is a subalgebra, but we will not even use this fact.]

Now let U be an open neighborhood of $0 \in G_e$ on which \exp is a diffeomorphism. Then $\exp(\mathfrak{h} \cap U)$ is a submanifold of G . It clearly suffices to show that if U is small enough, then

$$H \cap \exp(U) = \exp(\mathfrak{h} \cap U).$$

Choose a subspace $\mathfrak{h}' \subset G_e$ complementary to \mathfrak{h} , so that $G_e = \mathfrak{h} \oplus \mathfrak{h}'$.

Assertion 2. The map $\phi: G_e \rightarrow G$ defined by

$$\phi(X + X') = \exp X \exp X' \quad X \in \mathfrak{h}, X' \in \mathfrak{h}'$$

is a diffeomorphism in some neighborhood of 0.

Proof. Choose a basis $X_1, \dots, X_k, \dots, X_n$ of G_e with X_1, \dots, X_k a basis for \mathfrak{h} . Then ϕ is given by

$$\phi\left(\sum_{i=1}^n a_i X_i\right) = \exp\left(\sum_{i=1}^k a_i X_i\right) \exp\left(\sum_{i=k+1}^n a_i X_i\right).$$

Since the map $\sum_{i=1}^n a_i X_i \mapsto (a_1, \dots, a_n)$ is a diffeomorphism of G_e onto \mathbb{R}^n , it suffices to show that

$$\psi(a_1, \dots, a_n) = \exp\left(\sum_{i=1}^k a_i X_i\right) \exp\left(\sum_{i=k+1}^n a_i X_i\right)$$

is a diffeomorphism in a neighborhood of $0 \in \mathbb{R}^n$. This is clear, since

$$\psi_*\left(\frac{\partial}{\partial x^i}\bigg|_0\right) = X_i. \quad \text{Q.E.D.}$$

Assertion 3. There is a neighborhood V' of 0 in \mathfrak{h}' such that $\exp X' \notin H$ if $0 \neq X' \in V'$.

Proof. Choose an inner product on \mathfrak{h}' and let $K \subset \mathfrak{h}'$ be the compact set of all $X' \in \mathfrak{h}'$ with $1 \leq |X'| \leq 2$. If the assertion were false, there would be $X'_i \in \mathfrak{h}'$ with $X'_i \rightarrow 0$ and $\exp X'_i \in H$. Choose integers n_i with

$$n_i X'_i \in K.$$

Choosing a subsequence if necessary, we can assume $X'_i \rightarrow X' \in K$. Since

$$1/n_i \rightarrow 0, \quad \exp(1/n_i)(n_i X'_i) \in H,$$

it follows from *Assertion 1* that $X' \in \mathfrak{h}$, a contradiction. Q.E.D.

We can now complete the proof of the theorem. Choose a neighborhood $U = W \times W'$ of G_e on which \exp is a diffeomorphism, with

W a neighborhood of $0 \in \mathfrak{h}$

W' a neighborhood of $0 \in \mathfrak{h}'$

such that W' is contained in V' of *Assertion 3*, and ϕ of *Assertion 2* is a diffeomorphism on $W \times W'$. Clearly

$$\exp(\mathfrak{h} \cap U) \subset H \cap \exp(U).$$

To prove the reverse inclusion, let $a \in H \cap \exp(U)$. Then

$$a = \exp X \exp X' \quad X \in W, X' \in W'.$$

Since $a, \exp X \in H$ we obtain $\exp X' \in H$, so $0 = X'$, and $a \in \exp(\mathfrak{h} \cap U)$. ♦

Up to now, we have concentrated on the left invariant vector fields, but many properties of Lie groups are better expressed in terms of forms. A form ω is called left invariant if $L_a^*\omega = \omega$ for all $a \in G$. This means that

$$\omega(b) = L_a^*[\omega(ab)].$$

Clearly, a left invariant k -form ω is determined by its value $\omega(e) \in \Omega^k(G_e)$. Hence, if $\omega^1, \dots, \omega^n$ are left invariant 1-forms such that $\omega^1(e), \dots, \omega^n(e)$ span G_e^* , then every left invariant k -form is

$$\sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} \omega^{i_1} \wedge \dots \wedge \omega^{i_k} = \sum_I A_I \omega^I$$

for certain constants a_I . If $\omega^1(e), \dots, \omega^n(e)$ is the dual basis to $X_1, \dots, X_n \in G_e$, then any C^∞ vector field X can be written

$$X = \sum_{j=1}^n f^j \tilde{X}_j \quad \text{for } C^\infty \text{ functions } f^j.$$

Then

$$\omega^i(X) = f^i,$$

so ω^i is C^∞ . It follows that any left invariant form is C^∞ .

If ω is left invariant, then for $a \in G$ we have

$$L_a^*d\omega = d(L_a^*\omega) = d\omega,$$

so $d\omega$ is also left invariant. The formula on page 215 implies that for a left invariant 1-form ω and left invariant vector fields \tilde{X} and \tilde{Y} we have

$$\begin{aligned} d\omega(\tilde{X}, \tilde{Y}) &= \tilde{X}(\omega(\tilde{Y})) - \tilde{Y}(\omega(\tilde{X})) - \omega([\tilde{X}, \tilde{Y}]) \\ &= -\omega([\tilde{X}, \tilde{Y}]). \end{aligned}$$

Hence

$$(*) \quad d\omega(e)(X, Y) = -\omega(e)([X, Y]),$$

the bracket being the operation in \mathfrak{g} .

The interplay between left invariant and right invariant vector fields is the subject of Problem 11. Here we consider the case of forms.

16. PROPOSITION. Let $\psi: G \rightarrow G$ be $\psi(a) = a^{-1}$.

- (1) A form ω is left invariant if and only if $\psi^*\omega$ is right invariant.
- (2) If $\omega_e \in \Omega^k(G_e)$, then $\psi^*\omega_e = (-1)^k \omega_e$.
- (3) If ω is left and right invariant, then $d\omega = 0$.
- (4) If G is abelian, then \mathfrak{g} is abelian (converse of Corollary 7).

PROOF. (1) Clearly

$$\psi \circ R_b = L_{b^{-1}} \circ \psi,$$

so

$$R_b^* \psi^* = \psi^* L_{b^{-1}}^*.$$

If ω is left invariant, then

$$R_b^*(\psi^*\omega) = \psi^* L_{b^{-1}}^* \omega = \psi^*\omega.$$

so $\psi^*\omega$ is right invariant. The converse is similar.

(2) It clearly suffices to prove this for $k = 1$. So it is enough to show that $\psi_{*e}(X) = -X$ for $X \in G_e$. Now X is the tangent vector at $t = 0$ of the curve $t \mapsto \exp tX$. So $\psi_{*e}X$ is the tangent vector at $t = 0$ of $t \mapsto (\exp tX)^{-1} = \exp(-tX)$; this tangent vector is just $-X$.

(3) If ω is a left and right invariant k -form, then

$$\psi^*(\omega_e) = (-1)^k \omega_e.$$

Since $\psi^*\omega$ and ω are both left invariant, we have

$$\psi^*\omega = (-1)^k \omega.$$

The form $d\omega$ is also left and right invariant, so

$$\psi^*(d\omega) = (-1)^{k+1} d\omega.$$

But

$$\psi^*(d\omega) = d(\psi^*\omega) = d((-1)^k \omega) = (-1)^k d\omega.$$

So $d\omega = 0$.

(4) If G is abelian, then all left invariant 1-forms ω are also right invariant. So $d\omega = 0$ for all left invariant 1-forms. It follows from (*) that $[X, Y] = 0$ for all $X, Y \in \mathfrak{g}$.

Alternate proof of (4). By Theorem 14, if G is abelian, then for $X, Y \in G_e$ we have

$$\frac{t^2}{2}[X, Y] + O(t^3) = \frac{t^2}{2}[Y, X] + O(t^3).$$

Hence

$$\frac{1}{2}[X, Y] + O(t^3)/t^2 = \frac{1}{2}[Y, X] + O(t^3)/t^2.$$

Letting $t \rightarrow 0$, we obtain $[X, Y] = [Y, X]$. ♦

Since $d\omega$ is left invariant for any left invariant ω , it follows that for a basis $\omega^1, \dots, \omega^n$ of invariant 1-forms we can express each $d\omega^k$ in terms of the $\omega^i \wedge \omega^j$. First choose $X_1, \dots, X_n \in G_e$ dual to $\omega^1(e), \dots, \omega^n(e)$. There are constants C_{ij}^k such that

$$[X_i, X_j] = \sum_{k=1}^n C_{ij}^k X_k;$$

clearly we also have

$$[\tilde{X}_i, \tilde{X}_j] = \sum_{k=1}^n C_{ij}^k \tilde{X}_k.$$

The numbers C_{ij}^k are called the **constants of structure** of G (with respect to the basis X_1, \dots, X_n of \mathfrak{g}). From skew-symmetry of $[\ , \]$ and the Jacobi identity we obtain

- (1) $C_{ij}^k = -C_{ji}^k$
- (2) $\sum_{h=1}^n (C_{ij}^h C_{hk}^l + C_{jk}^h C_{hi}^l + C_{ki}^h C_{hj}^l) = 0.$

From (*) on page 394 we obtain

$$d\omega^k = - \sum_{i < j} C_{ij}^k \omega^i \wedge \omega^j = - \frac{1}{2} \sum_{i, j} C_{ij}^k \omega^i \wedge \omega^j.$$

It turns out that (2) is exactly what we obtain from the relation $d^2\omega^k = 0$. Condition (2) is thus an integrability condition. In fact, we can prove (Problem 30) that if C_{ij}^k are constants satisfying (1) and (2), then we can find everywhere linearly independent 1-forms $\omega^1, \dots, \omega^n$ in a neighborhood of $0 \in \mathbb{R}^n$ such that

$$d\omega^k = - \frac{1}{2} \sum_{i, j} C_{ij}^k \omega^i \wedge \omega^j.$$

Moreover, the existence of such ω^i implies (Problem 29) that we can define a multiplication $(a, b) \mapsto ab$ in a neighborhood of 0 which is a group as far as it can be and which has the ω^i as left invariant 1-forms. From this latter fact and (a suitable local version of) Theorem 5 we could immediately deduce the following Theorem, for which we supply an independent proof.

17. THEOREM. Let G be a Lie group with a basis of left invariant 1-forms $\omega^1, \dots, \omega^n$ and constants of structure C_{ij}^k . Let M^n be a differentiable manifold and let $\theta^1, \dots, \theta^n$ be everywhere linearly independent 1-forms on M satisfying

$$d\theta^k = - \sum_{i < j} C_{ij}^k \theta^i \wedge \theta^j.$$

Then for every $p \in M$ there is a neighborhood U and a diffeomorphism $f: U \rightarrow G$ such that

$$\theta^i = f^* \omega^i.$$

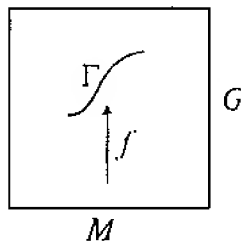
PROOF. Let $\pi_1: M \times G \rightarrow M$ and $\pi_2: M \times G \rightarrow G$ be the projections. Let

$$\bar{\theta}^k = \pi_1^* \theta^k, \quad \bar{\omega}^k = \pi_2^* \omega^k.$$

Then

$$\begin{aligned} d(\bar{\theta}^k - \bar{\omega}^k) &= - \sum_{i < j} C_{ij}^k ([\bar{\theta}^i \wedge \bar{\theta}^j] - [\bar{\omega}^i \wedge \bar{\omega}^j]) \\ &= - \sum_{i < j} C_{ij}^k [\bar{\theta}^i \wedge (\bar{\theta}^j - \bar{\omega}^j) + (\bar{\theta}^i - \bar{\omega}^i) \wedge \bar{\omega}^j]. \end{aligned}$$

By Proposition 7-14, $M \times G$ is foliated by n -dimensional manifolds whose tangent spaces at each point are annihilated by all $\bar{\theta}^k - \bar{\omega}^k$. Choose $a \in G$ and let Γ be the folium through (p, a) . Now $\bar{\theta}^1, \dots, \bar{\theta}^n, \bar{\omega}^1, \dots, \bar{\omega}^n$ are linearly independent everywhere; so on $\Gamma_{(p,a)}$, which is the set of vectors in $(M \times G)_{(p,a)}$ where $\bar{\theta}^k - \bar{\omega}^k = 0$, the sets $\bar{\theta}^1, \dots, \bar{\theta}^n$ and $\bar{\omega}^1, \dots, \bar{\omega}^n$ are each linearly independent. Hence $\pi_1: \Gamma \rightarrow M$ and $\pi_2: \Gamma \rightarrow G$ are each diffeomorphisms in some neighborhood of (p, a) . This means that Γ contains the graph of a diffeomorphism f from a neighborhood U of p to a neighborhood of a .



Let $\tilde{f}: U \rightarrow M \times G$ be the map

$$\tilde{f}(q) = (q, f(q)) \in \Gamma.$$

Since $\tilde{\theta}^k - \tilde{\omega}^k = 0$ on Γ , we have

$$\begin{aligned} 0 &= \tilde{f}^*(\tilde{\theta}^k - \tilde{\omega}^k) = \tilde{f}^*\pi_1^*\theta^k - \tilde{f}^*\pi_2^*\omega^k \\ &= (\pi_1 \circ \tilde{f})^*\theta^k - (\pi_2 \circ \tilde{f})^*\omega^k \\ &= \theta^k - f^*\omega^k. \quad \blacklozenge \end{aligned}$$

It is also possible to say by how much any two such maps differ:

18. THEOREM. Let M be a connected manifold, let G be a Lie group, and let $f_1, f_2: M \rightarrow G$ be two C^∞ maps such that

$$f_1^*(\omega) = f_2^*(\omega)$$

for all left invariant 1-forms ω . Then f_1 and f_2 differ by a left translation, that is, there is a (unique) $a \in G$ such that

$$f_2 = L_a \circ f_1.$$

PEDESTRIAN PROOF. Case 1. $M = \mathbb{R}$ and the two maps $\gamma_1, \gamma_2: \mathbb{R} \rightarrow G$ satisfy $\gamma_1(0) = \gamma_2(0)$. We must show that $\gamma_1 = \gamma_2$. For every left invariant 1-form ω we have

$$\begin{aligned} \omega(\gamma_2(t)) \left(\frac{d\gamma_2}{dt} \right) &= \gamma_2^*\omega \left(\frac{d}{dt} \Big|_t \right) = \gamma_1^*\omega \left(\frac{d}{dt} \Big|_t \right) \\ &= \omega(\gamma_1(t)) \left(\frac{d\gamma_1}{dt} \right) \\ &= \left[(L_{\gamma_2(t)\gamma_1(t)^{-1}})^*\omega(\gamma_2(t)) \right] \left(\frac{d\gamma_1}{dt} \right) \\ &= \omega(\gamma_2(t)) \left(\left[L_{\gamma_2(t)\gamma_1(t)^{-1}} \right]_* \frac{d\gamma_1}{dt} \right). \end{aligned}$$

It follows that

$$\frac{d\gamma_2}{dt} = \left[L_{\gamma_2(t)\gamma_1(t)^{-1}} \right]_* \frac{d\gamma_1}{dt}.$$

If we regard γ_1 as given, and write this equation out in a coordinate system, then it becomes an ordinary differential equation for γ_2 (of the type considered

in the Addendum to Chapter 5), so it has a unique solution with the initial condition $\gamma_2(0) = \gamma_1(0)$. But this solution is clearly $\gamma_2 = \gamma_1$.

Case 2. $M = \mathbb{R}$, but the maps γ_1, γ_2 are arbitrary. Choose $a \in G$ so that

$$\gamma_2(0) = a \cdot \gamma_1(0).$$

If ω is a left invariant 1-form, then

$$(L_a \circ \gamma_1)^*(\omega) = \gamma_1^*(L_a^*\omega) = \gamma_1^*(\omega) = \gamma_2^*(\omega).$$

Since $L_a \circ \gamma_1(0) = \gamma_2(0)$, it follows from *Case 1* that $L_a \circ \gamma_1 = \gamma_2$.

Case 3. General case. Let $p_0 \in M$. Choose $a \in G$ so that

$$f_2(p_0) = a \cdot f_1(p_0).$$

For any $p \in M$ there is a C^∞ curve $c: \mathbb{R} \rightarrow M$ with $c(0) = p_0$ and $c(1) = p$. Let $\gamma_i = f_i \circ c$. Then

$$\gamma_2^*(\omega) = c^* f_2^*(\omega) = c^* f_1^*(\omega) = \gamma_1^*(\omega).$$

By *Case 2*, we have

$$\gamma_2(t) = a \cdot \gamma_1(t) \quad \text{for all } t.$$

in particular for $t = 1$, so $f_2(p) = a \cdot f_1(p)$.

ELEGANT PROOF. Let $\pi_i: G \times G \rightarrow G$ be projection on the i^{th} factor. Choose a basis $\omega^1, \dots, \omega^n$ for the left invariant 1-forms. For $(a, b) \in G \times G$, let

$$\Delta_{(a,b)} = \bigcap_{i=1}^n \ker(\pi_1^* \omega^i - \pi_2^* \omega^i).$$

Then Δ is an integrable distribution on $G \times G$. In fact, if $\Delta(G) \subset G \times G$ is the diagonal subgroup $\{(a, a) : a \in G\}$, then the maximal integral manifolds of Δ are the left cosets of $\Delta(G)$. Now define $h: M \rightarrow G \times G$ by

$$h(p) = (f_1(p), f_2(p)).$$

By assumption,

$$h^*(\pi_1^* \omega^i - \pi_2^* \omega^i) = f_1^* \omega^i - f_2^* \omega^i = 0.$$

Since M is connected, it follows that $h(M)$ is contained in some left coset of $\Delta(G)$. In other words, there are $a, b \in G$ with

$$af_1(p) = bf_2(p) \quad \text{for all } p \in M. \quad \spadesuit$$

19. COROLLARY. If G is a connected Lie group and $f: G \rightarrow G$ is a C^∞ map preserving left invariant forms, then $f = L_a$ for a unique $a \in G$.

While left invariant 1-forms play a fundamental role in the study of G , the left invariant n -forms are also very important. Clearly, all left invariant n -forms are a constant multiple of any non-zero one. If σ^n is a left invariant n -form, then σ^n determines an orientation on G , and if $f: G \rightarrow \mathbb{R}$ is a C^∞ function with compact support, we can define

$$\int_G f \sigma^n.$$

Since σ^n is usually kept fixed in any discussion, this is often abbreviated to

$$\int_G f \quad \text{or} \quad \int_G f(a) da.$$

The latter notation has advantages in certain cases. For example, left invariance of σ^n implies that

$$\int_G f(a) da = \int_G f(ba) da.$$

in other words,

$$\int_G f \sigma^n = \int_G g \sigma^n, \quad \text{where } g(a) = f(ba);$$

[note that L_b is an orientation preserving diffeomorphism, so

$$\int_G f \sigma^n = \int_G L_b^*(f \sigma^n) = \int_G (f \circ L_b) L_b^* \sigma^n = \int_G (f \circ L_b) \sigma^n,$$

which proves the formula]. We can, of course, also consider right invariant n -forms. These generally turn out to be quite different from the left invariant n -forms (see the example in Problem 25). But in one case they coincide.

20. PROPOSITION. If G is compact and connected and ω is a left invariant n -form, then ω is also right invariant.

PROOF. Suppose $\omega \neq 0$. For each $a \in G$, the form $R_a^* \omega$ is left invariant, so there is a unique real number $f(a)$ with

$$R_a^* \omega = f(a) \omega.$$

Since $R_a^* \circ R_b^* = (R_{ab})^*$, we have

$$f(ab) = f(ba) = f(a) \cdot f(b).$$

So $f(G) \subset \mathbb{R}$ is a compact connected subgroup of $\mathbb{R} - \{0\}$. Hence $f(G) = \{1\}$. ♦

We can also consider Riemannian metrics on G . In the case of a compact group G there is always a Riemannian metric on G which is both left and right invariant. In fact, if (\cdot, \cdot) is any Riemannian metric we can choose a bi-invariant n -form σ^n and define a bi-invariant $\langle\langle \cdot, \cdot \rangle\rangle$ on G by

$$\langle\langle V, W \rangle\rangle = \int_{G \times G} (L_{a*} R_{b*}(V), L_{a*} R_{b*}(W)) da db.$$

We are finally ready to account for some terminology from Chapter 9.

21. PROPOSITION. Let G be a Lie group with a bi-invariant metric.

- (1) For any $a \in G$, the map $I_a: G \rightarrow G$ given by $I_a(b) = ab^{-1}a$ is an isometry which reverses geodesics through a , i.e., if γ is a geodesic and $\gamma(0) = a$, then $I_a(\gamma(t)) = \gamma(-t)$.
- (2) The geodesics γ with $\gamma(0) = e$ are precisely the 1-parameter subgroups of G , i.e., the maps $t \mapsto \exp(tX)$ for some $X \in \mathfrak{g}$.

PROOF. (1) Since

$$I_e(b) = b^{-1},$$

the map $I_{e*}: G_e \rightarrow G_e$ is just multiplication by -1 (see the proof of Proposition 16(2)), so it is an isometry on G_e . Since

$$I_e = R_{a^{-1}} I_e L_{a^{-1}}$$

for any $a \in G$, the map $I_{e*}: G_a \rightarrow G_{a^{-1}}$ is also an isometry. Clearly I_e reverses geodesics through e .

Since

$$I_a = R_a I_e R_a^{-1},$$

it is clear that I_a is an isometry reversing geodesic through a .

- (2) Let $\gamma: \mathbb{R} \rightarrow G$ be a geodesic with $\gamma(0) = e$. For fixed t , let

$$\tilde{\gamma}(u) = \gamma(t + u).$$

Then $\tilde{\gamma}$ is a geodesic and $\tilde{\gamma}(0) = \gamma(t)$. So

$$\begin{aligned} I_{\gamma(t)} I_e(\gamma(u)) &= I_{\gamma(t)}(\gamma(-u)) = I_{\gamma(t)}(\tilde{\gamma}(-u - t)) \\ &= \tilde{\gamma}(t + u) = \gamma(u + 2t). \end{aligned}$$

But also

$$I_{\gamma(t)} I_e(b) = \gamma(t)b\gamma(t),$$

so

$$\gamma(t)\gamma(u)\gamma(t) = \gamma(u + 2t).$$

It follows by induction that

$$\gamma(nt) = \gamma(t)^n \quad \text{for any integer } n.$$

If $t' = n't$ and $t'' = n''t$ for integers n' and n'' , then

$$\gamma(t' + t'') = \gamma(t)^{n' + n''} = \gamma(t')\gamma(t''),$$

so γ is a homomorphism on \mathbb{Q} . By continuity, γ is a 1-parameter subgroup.

These are the only geodesics, since there are 1-parameter subgroups with any tangent vector at $t = 0$, and geodesics through e are determined by their tangent vectors at $t = 0$. ♦

We conclude this chapter by introducing some neat formalism which allows us to write the expression for $d\omega^k$ in an invariant way that does not use the constants of structure of G . If V is a d -dimensional vector space, we define a V -valued k -form on M to be a function ω such that each $\omega(p)$ is an alternating map

$$\omega(p): \underbrace{M_p \times \cdots \times M_p}_{k \text{ times}} \rightarrow V.$$

If v_1, \dots, v_d is a basis for V , then there are ordinary k -forms $\omega^1, \dots, \omega^d$ such that for $X_1, \dots, X_k \in M_p$ we have

$$\omega(p)(X_1, \dots, X_k) = \sum_{i=1}^d \omega^i(p)(X_1, \dots, X_k) v_i;$$

we will write simply

$$\omega = \sum_{i=1}^d \omega^i \cdot v_i.$$

For any V -valued k -form ω we define a V -valued $(k+1)$ -form $d\omega$ by

$$d\omega = \sum_{i=1}^d d\omega^i \cdot v_i;$$

a simple calculation shows that this definition does not depend on the choice of basis v_1, \dots, v_d for V .

Similarly, suppose $\rho: U \times V \rightarrow W$ is a bilinear map, where U and V have bases u_1, \dots, u_c and v_1, \dots, v_d , respectively. If ω is a U -valued k -form

$$\omega = \sum_{i=1}^c \omega^i \cdot u_i$$

and η is a V -valued l -form

$$\eta = \sum_{j=1}^d \eta^j \cdot v_j,$$

then

$$\sum_{i=1}^c \sum_{j=1}^d \omega^i \wedge \eta^j \cdot \rho(u_i, v_j)$$

is a W -valued $(k+l)$ -form; a calculation shows that this does not depend on the choice of bases u_1, \dots, u_c or v_1, \dots, v_d . We will denote this W -valued $(k+l)$ -form by $\rho(\omega \wedge \eta)$.

These concepts have a natural place in the study of a Lie group G . Although there is no natural way to choose a basis of left invariant 1-forms on G , there is a natural \mathfrak{g} -valued 1-form on G , namely the form ω defined by

$$(*) \quad \omega(a)(\tilde{X}(a)) = X \in \mathfrak{g}.$$

Using the bilinear map $[\cdot, \cdot]: \mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$, we have, for any \mathfrak{g} -valued k -form η and any \mathfrak{g} -valued l -form λ on G , a new \mathfrak{g} -valued $(k+l)$ -form $[\eta \wedge \lambda]$ on G .

Now suppose that $X_1, \dots, X_n \in G_e = \mathfrak{g}$ is a basis, and that $\omega^1, \dots, \omega^n$ is a dual basis of left invariant 1-forms. The form ω defined by $(*)$ can clearly be written

$$\omega = \sum_{k=1}^n \omega^k \cdot X_k.$$

Then

$$\begin{aligned} (I) \quad d\omega &= \sum_{k=1}^n d\omega^k \cdot X_k \\ &= \sum_{k=1}^n \left(\sum_{i < j} C_{ij}^k \omega^i \wedge \omega^j \right) \cdot X_k. \end{aligned}$$

On the other hand,

$$[X_i, X_j] = \sum_{k=1}^n C_{ij}^k X_k,$$

so

$$(2) \quad [\omega \wedge \omega] = \sum_{k=1}^n \left(\sum_{i=1}^n \sum_{j=1}^n C_{ij}^k \omega^i \wedge \omega^j \cdot X_k \right).$$

Comparing (1) and (2), we obtain the equations of structure of G :

$$d\omega = -\frac{1}{2}[\omega \wedge \omega].$$

The equations of structure of a Lie group will play an important role in Volume III. For the present we merely wish to point out that the terms $d\omega$ and $[\omega \wedge \omega]$ appearing in this equation can also be defined in an invariant way. For the term $d\omega$ we just modify the formula in Theorem 7-13: If U is a vector field on G and f is a \mathfrak{g} -valued function on G , then (Problem 20) we can define a \mathfrak{g} -valued function $U(f)$ on G . On the other hand, $\omega(U)$ is a \mathfrak{g} -valued function on G . For vector fields U and V we can then define

$$d\omega(U, V) = U(\omega(V)) - V(\omega(U)) - \omega([U, V]).$$

Recall that the value at $a \in G$ of the right side depends only on the values U_a and V_a of U and V at a . If we choose $U = \tilde{X}$, $V = \tilde{Y}$ for some $X, Y \in G_e$, then

$$\begin{aligned} d\omega(a)(\tilde{X}_a, \tilde{Y}_a) &= 0 - 0 - \omega(a)([\tilde{X}, \tilde{Y}]_a) \\ &= -\omega(e)([\tilde{X}, \tilde{Y}]_e) && \text{since } [\tilde{X}, \tilde{Y}] \text{ is left invariant} \\ &= -\omega(e)([X, Y]) && \text{by definition of } [\cdot, \cdot] \text{ in } G_e \\ &= -[X, Y] \\ &= -[\omega(a)(\tilde{X}_a), \omega(a)(\tilde{Y}_a)] \end{aligned} \quad \left. \vphantom{\begin{aligned} d\omega(a)(\tilde{X}_a, \tilde{Y}_a) &= 0 - 0 - \omega(a)([\tilde{X}, \tilde{Y}]_a) \\ &= -\omega(e)([\tilde{X}, \tilde{Y}]_e) \\ &= -\omega(e)([X, Y]) \\ &= -[X, Y] \end{aligned}} \right\} \text{by definition of } \omega.$$

It follows that for any vector fields U and V we have

$$d\omega(U, V) = -[\omega(U), \omega(V)].$$

Problem 20 gives an invariant definition of $\rho(\omega \wedge \eta)$ and shows that this equation is equivalent to the equations of structure.

WARNING: In some books the equation which we have just deduced appears as $d\omega(U, V) = -\frac{1}{2}[\omega(U), \omega(V)]$. The appearance of the factor $\frac{1}{2}$ here has *nothing* to do with the $\frac{1}{2}$ in the other form of the structure equations. It comes about because some books do not use the factor $(k+l)!/k!l!$ in the definition of \wedge . This makes their $\lambda \wedge \eta$ equal to $\frac{1}{2}$ of ours for 1-forms λ and η . Then the definition of $d(\sum \omega_i dx^i)$ as $\sum d\omega_i \wedge dx^i$ makes their $d\omega$ equal to $\frac{1}{2}$ of ours for 1-forms ω .

PROBLEMS

1. Let G be a group which is also a C^∞ manifold, and suppose that $(x, y) \mapsto xy$ is C^∞ .

(a) Find f^{-1} when $f: G \times G \rightarrow G \times G$ is $f(x, y) = (x, xy)$.

(b) Show that (e, e) is a regular point of f .

(c) Conclude that G is a Lie group.

2. Let G be a topological group, and $H \subset G$ a subgroup. Show that the closure \bar{H} of H is also a subgroup.

3. Let G be a topological group and $H \subset G$ a subgroup.

(a) If H is open, then so is every coset gH .

(b) If H is open, then H is closed.

4. Let G be a connected topological group, and U a neighborhood of $e \in G$. Let U^n denote all products $a_1 \cdots a_n$ for $a_i \in U$.

(a) Show that U^{n+1} is a neighborhood of U^n .

(b) Conclude that $\bigcup_n U^n = G$. (Use Problem 3.)

(c) If G is locally compact and connected, then G is σ -compact.

5. Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be distance preserving, with $f(0) = 0$.

(a) Show that f takes straight lines to straight lines.

(b) Show that f takes planes to planes.

(c) Show that f is a linear transformation, and hence an element of $O(n)$.

(d) Show that any element of $E(n)$ can be written $A \cdot \tau$ for $A \in O(n)$ and τ a translation.

6. Show that the tangent bundle TG of a Lie group G can always be made into a Lie group.

7. We have computed that for $M \in \mathfrak{gl}(n, \mathbb{R})$ we have

$$\tilde{M} = \sum_{k,l} \tilde{M} x^{kl} \cdot \frac{\partial}{\partial x^{kl}}, \quad \text{where } \tilde{M} x^{kl}(A) = \sum_{\alpha=1}^n M_{\alpha l} A_{k\alpha}.$$

(a) Show that this means that

$$\tilde{M}(A) = A \cdot M \in \mathbb{R}^{n^2} = \text{GL}(n, \mathbb{R})_A.$$

(It is actually clear *a priori* that \tilde{M} defined in this way is left invariant, for $L_{A*} = L_A$ since L_A is linear.)

(b) Find the right invariant vector field with value M at I .

8. Let G and H be topological groups and $\phi: U \rightarrow H$ a map on a connected open neighborhood U of $e \in G$ such that $\phi(ab) = \phi(a)\phi(b)$ when $a, b, ab \in U$.

(a) For each $c \in G$, consider pairs (V, ψ) , where $V \subset G$ is an open neighborhood of c with $V \cdot V^{-1} \subset U$, and where $\psi: V \rightarrow H$ satisfies $\psi(a) \cdot \psi(b)^{-1} = \phi(ab^{-1})$ for $a, b \in V$. Define $(V_1, \psi_1) \sim (V_2, \psi_2)$ if $\psi_1 = \psi_2$ on some smaller neighborhood of c . Show that the set of all \sim equivalence classes, for all $c \in G$, can be made into a covering space of G .

(b) Conclude that if G is simply-connected, then ϕ can be extended uniquely to a homomorphism of G into H .

9. In Theorem 5, show that ϕ and ψ are equal even if they are defined only on a neighborhood U of $e \in G$, provided that U is connected.

10. Show that Corollary 7 is false if G is not assumed connected.

11. If G is a group, we define the **opposite group** G° to be the same set with the multiplication \star defined by $a \star b = b \cdot a$. If \mathfrak{g} is a Lie algebra, with operation $[\ , \]$, we define the **opposite Lie algebra** \mathfrak{g}° to be the same set with the operation $[X, Y]^\circ = -[X, Y]$.

(a) G° is a group, and if $\psi: G \rightarrow G$ is $a \mapsto a^{-1}$, then ψ is an isomorphism from G to G° .

(b) \mathfrak{g}° is a Lie algebra, and $X \mapsto -X$ is an isomorphism of \mathfrak{g} onto \mathfrak{g}° .

(c) $\mathcal{L}(G^\circ)$ is isomorphic to $[\mathcal{L}(G)]^\circ = \mathfrak{g}^\circ$.

(d) Let $[\ , \]$ be the operation on G_e obtained by using right invariant vector fields instead of left invariant ones. Then $(\mathfrak{g}, [\ , \])$ is isomorphic to $\mathcal{L}(G^\circ)$, and hence to \mathfrak{g}° .

(e) Use this to give another proof that \mathfrak{g} is abelian when G is abelian.

12. (a) Show that

$$\exp \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} = \begin{pmatrix} \cos a & \sin a \\ -\sin a & \cos a \end{pmatrix}.$$

(b) Use the matrices A and B below to show that $\exp(A + B)$ is not generally equal to $(\exp A)(\exp B)$.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}$$

13. Let $X, Y \in G_e$ with $[X, Y] = 0$.

(a) Use Lemma 5-13 to show that $(\exp sX)(\exp tY) = (\exp tY)(\exp sX)$.

(b) More generally, use Theorem 5 to show that \exp is a homomorphism on the subspace of G_e spanned by X and Y . In particular, $\exp(X + Y) = (\exp X)(\exp Y)$.

14. Problem 13 implies that $\exp t(X + Y) = (\exp tX)(\exp tY)$ if $[X, Y] = 0$. A more general result holds. Let X and Y be vector fields on a C^∞ manifold M with corresponding local 1-parameter families of local diffeomorphisms $\{\phi_t\}$, $\{\psi_s\}$. Suppose that $[X, Y] = 0$, and let $\eta_t = \phi_t \circ \psi_t = \psi_t \circ \phi_t$.

(a) Show that

$$\frac{d\eta_t(p)}{dt} = X(\eta_t(p)) + \phi_{t*}(Y(\psi_t(p))).$$

(b) Using Corollary 5-12, show that

$$\frac{d\eta_t(p)}{dt} = X(\eta_t(p)) + Y(\eta_t(p)).$$

In other words, $\{\eta_t\}$ is generated by $X + Y$.

15. (a) If M is a diagonal matrix with complex entries, show that

$$\det \exp M = e^{\text{trace } M}.$$

(b) Show that the same equation holds for all diagonalizable M with complex entries.

(c) Conclude that it holds for all M with complex entries. (The diagonalizable matrices are dense; compare Problem 7-15.)

(d) Using Proposition 9, show that for the homomorphism $\det: \text{GL}(n, \mathbb{R}) \rightarrow \mathbb{R} - \{0\}$, the map $\det_*: \mathfrak{gl}(n, \mathbb{R}) \rightarrow \mathcal{L}(\mathbb{R} - \{0\}) = \mathbb{R}$ is just $M \mapsto \text{trace } M$.

(e) Use this fact to give a fancy proof that $\text{trace } MN = \text{trace } NM$. (Look at $\text{trace}(MN - NM) = \text{trace}[M, N]$.)

(f) Prove the result in part (d) directly, without using (c). (Since \det_* and trace are homomorphisms, it suffices to look at matrices with only one non-zero entry.)

(g) Now use this result and Proposition 9 to give a fancy proof of (c).

16. (a) Let U be a neighborhood of the identity $(1, 0)$ of S^1 (considered as a subset of \mathbb{R}^2). Show that no matter how small U is, there are elements $a \in U$ which have square roots outside U in addition to their square root in U .

(b) Show that for each $n \geq 1$, there is a neighborhood U of $e \in G$ such that every element in U has a unique n^{th} root in U .

(c) For $G = S^1$, show that there is no neighborhood U which has this property for all n .

17. (a) Let (x, V) be a coordinate system around $e \in G$ with $x^i(e) = 0$. Let

$$x^i(ab) = f^i(x^1(a), \dots, x^n(a), x^1(b), \dots, x^n(b))$$

for C^∞ functions f^i . Show that

$$D_j f^i(0) = D_{n+j} f^i(0) = \delta_j^i.$$

(b) If $\alpha, \beta: (-\varepsilon, \varepsilon) \rightarrow G$ are differentiable, show that

$$(\alpha \cdot \beta)'(0) = \alpha'(0) + \beta'(0).$$

(c) Also deduce this result from Theorem 14(1). (Not even the full strength of (1) is needed; it suffices to know that $\exp tX \exp tY = \exp\{t(X+Y) + O(t)\}$. The argument of part (a) is essentially equivalent to the initial part of the deduction of (1).)

18. Let G be a Lie group, and let $H \subset G$ be a subgroup of G (algebraically), such that every $a \in H$ can be joined to e by a C^∞ path lying in H . Let $\mathfrak{h} \subset G_e$ be the set of tangent vectors to all C^∞ paths lying in H .

(a) Show that \mathfrak{h} is a subalgebra of G_e . (Use Theorem 14.)

(b) Let $K \subset G$ be the connected Lie subgroup of G with Lie algebra \mathfrak{h} . Show that $H \subset K$. *Hint:* Join any $a \in H$ to e by a C^∞ curve c , and show that the tangent vectors of c lie in the distribution constructed in the proof of Theorem 4.

(c) Let c_1, \dots, c_k be curves in H with $\{c_i'(0)\}$ a basis for \mathfrak{h} . By considering the map $f(t^1, \dots, t^k) = c_1(t^1) \cdots c_k(t^k)$, show that $K \subset H$. Thus, H is a Lie subgroup of G . It is even true that $H \subset G$ is a Lie subgroup if H is path connected (by not necessarily C^∞ paths); see Yamabe, *On an arcwise connected subgroup of a Lie group*, Osaka Math. J. 2 (1950), 13–14.

(d) If $H \subset G$ is a subgroup and an immersed submanifold, then H is a Lie subgroup.

19. For $a \in G$, consider the map $b \mapsto aba^{-1} = L_a R_a^{-1}(b)$. The map

$$(L_a R_a^{-1})_*: \mathfrak{g} \rightarrow \mathfrak{g}$$

is denoted by $\text{Ad}(a)$; usually $\text{Ad}(a)(X)$ is denoted simply by $\text{Ad}(a)X$.

(a) $\text{Ad}(ab) = \text{Ad}(a) \circ \text{Ad}(b)$. Thus we have a homomorphism $\text{Ad}: G \rightarrow \text{Aut}(\mathfrak{g})$, where $\text{Aut}(\mathfrak{g})$, the automorphism group of \mathfrak{g} , is the set of all non-singular linear transformations of the vector space \mathfrak{g} onto itself (thus, isomorphic to $\text{GL}(n, \mathbb{R})$ if \mathfrak{g} has dimension n). The map Ad is called the **adjoint representation**.

(b) Show that

$$\exp(\text{Ad}(a)X) = a(\exp X)a^{-1}.$$

Hint: This follows immediately from one of our propositions.

(c) For $A \in \mathrm{GL}(n, \mathbb{R})$ and $M \in \mathfrak{gl}(n, \mathbb{R})$ show that

$$\mathrm{Ad}(A)M = AMA^{-1}.$$

(It suffices to show this for M in a neighborhood of 0.)

(d) Show that

$$\mathrm{Ad}(\exp tX)Y = Y + t[X, Y] + O(t^2).$$

(e) Since $\mathrm{Ad}: G \rightarrow \mathfrak{g}$, we have the map

$$\mathrm{Ad}_{*e}: \mathfrak{g} (= G_e) \rightarrow \begin{array}{l} \text{tangent space of } \mathrm{Aut}(\mathfrak{g}) \text{ at the} \\ \text{identity map } 1_{\mathfrak{g}} \text{ of } \mathfrak{g} \text{ to itself.} \end{array}$$

This tangent space is isomorphic to $\mathrm{End}(\mathfrak{g})$, where $\mathrm{End}(\mathfrak{g})$ is the vector space of all linear transformations of \mathfrak{g} into itself. If c is a curve in $\mathrm{Aut}(\mathfrak{g})$ with $c(0) = 1_{\mathfrak{g}}$, then to regard $c'(0)$ as an element of $\mathrm{Aut}(\mathfrak{g})$, we let it operate on $Y \in \mathfrak{g}$ by

$$c'(0)(Y) = \left. \frac{d}{dt} \right|_{t=0} c(Y).$$

(Compare with the case $\mathfrak{g} = \mathbb{R}^n$, $\mathrm{Aut}(\mathfrak{g}) = \mathrm{GL}(n, \mathbb{R})$, $\mathrm{End}(\mathfrak{g}) = n \times n$ matrices.) Use (d) to show that

$$\mathrm{Ad}_{*e}(X)(Y) = [X, Y].$$

(A proof may also be given using the fact that $[\tilde{X}, \tilde{Y}] = L_{\tilde{X}}\tilde{Y}$.) The map $Y \mapsto [X, Y]$ is denoted by $\mathrm{ad} X \in \mathrm{End}(\mathfrak{g})$.

(f) Conclude that

$$\mathrm{Ad}(\exp X) = \exp(\mathrm{ad} X) = 1_{\mathfrak{g}} + \mathrm{ad} X + \frac{(\mathrm{ad} X)^2}{2!} + \cdots.$$

(g) Let G be a connected Lie group and $H \subset G$ a Lie subgroup. Show that H is a normal subgroup of G if and only if $\mathfrak{h} = \mathcal{L}(H)$ is an ideal of $\mathfrak{g} = \mathcal{L}(G)$, that is, if and only if $[X, Y] \in \mathfrak{h}$ for all $X \in \mathfrak{g}$, $Y \in \mathfrak{h}$.

20. (a) Let $f: M \rightarrow V$, where V is a finite dimensional vector space, with basis v_1, \dots, v_d . For $X_p \in M_p$, define $X_p(f) \in V$ by

$$X(f) = \sum_{i=1}^d X_p(f^i) \cdot v_i,$$

where $f = \sum_{i=1}^d f^i \cdot v_i$ for $f^i: M \rightarrow \mathbb{R}$. Show that this definition is independent of the choice of basis v_1, \dots, v_d for V .

- (b) If ω is a V -valued k -form, show that $d\omega$ may be defined invariantly by the formula in Theorem 7-13 (using the definition in part (a)).
- (c) For $\rho: U \times V \rightarrow W$, show that $\rho(\omega \wedge \eta)$ may be defined invariantly by

$$\begin{aligned} \rho(\omega \wedge \eta)(X_1, \dots, X_k, X_{k+1}, \dots, X_{k+l}) \\ = \frac{1}{k!l!} \sum_{\sigma \in S_{k+l}} \text{sgn } \sigma \cdot \rho(\omega(X_{\sigma(1)}, \dots, X_{\sigma(k)}), \eta(X_{\sigma(k+1)}, \dots, X_{\sigma(k+l)})). \end{aligned}$$

Conclude, in particular, that

$$[\omega \wedge \omega](X, Y) = 2[\omega(X), \omega(Y)].$$

- (d) Deduce the structure equations from (b) and (c).

21. (a) If ω is a U -valued k -form and η is a V -valued l -form, and $\rho: U \times V \rightarrow W$, then

$$d(\rho(\omega \wedge \eta)) = \rho(d\omega \wedge \eta) + (-1)^k \rho(\omega \wedge d\eta).$$

- (b) For a \mathfrak{g} -valued k -form ω and l -form η we have

$$[\omega \wedge \eta] = (-1)^{kl+1} [\eta \wedge \omega].$$

- (c) Moreover, if λ is a \mathfrak{g} -valued m -form, then

$$(-1)^{km} [\omega \wedge [\eta \wedge \lambda]] + (-1)^{kl} [\eta \wedge [\lambda \wedge \omega]] + (-1)^{lm} [\lambda \wedge [\eta \wedge \omega]] = 0.$$

22. Let $G \subset GL(n, \mathbb{R})$ be a Lie subgroup. The inclusion map $G \rightarrow GL(n, \mathbb{R}) \rightarrow \mathbb{R}^{n^2}$ will be denoted by P (for "point"). Then dP is an \mathbb{R}^{n^2} -valued 1-form (it corresponds to the identity map of the tangent space of G into itself). We can also consider dP as a matrix of 1-forms; it is just the matrix (dx^{ij}) , where each dx^{ij} is restricted to the tangent bundle of G . We also have the \mathbb{R}^{n^2} -valued 1-form (or matrix of 1-forms) $P^{-1} \cdot dP$, where \cdot denotes matrix multiplication, and P^{-1} denotes the map $A \mapsto A^{-1}$ on G .

- (a) $P^{-1} \cdot dP = \rho(P^{-1} \wedge dP)$, where $\rho: \mathbb{R}^{n^2} \times \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n^2}$ is matrix multiplication.
- (b) $L_A^* dP = A \cdot dP$. (Use $f^*d = df^*$.)
- (c) $P^{-1} \cdot dP$ is left invariant; and $(dP) \cdot P^{-1}$ is right invariant.
- (d) $P^{-1} \cdot dP$ is the natural \mathfrak{g} -valued 1-form ω on G . (It suffices to check that $P^{-1} \cdot dP = \omega$ at I .)
- (e) Using $dP = P \cdot \omega$, show that $0 = dP \cdot \omega + P \cdot d\omega$, where the matrix of 2-forms $P \cdot d\omega$ is computed by formally multiplying the matrices of 1-forms dP and ω . Deduce that

$$d\omega + \omega \cdot \omega = 0.$$

If ω is the matrix of 1-forms $\omega = (\omega^{ij})$, this says that

$$d\omega^{ij} = - \sum_k \omega^{ik} \wedge \omega^{kj}.$$

Check that these equations are equivalent to the equations of structure (use the form $d\omega(X, Y) = -[\omega(X), \omega(Y)]$.)

23. Let $G \subset \text{GL}(2, \mathbb{R})$ consist of all matrices $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ with $a \neq 0$. For convenience, denote the coordinates x^{11} and x^{12} on $\text{GL}(2, \mathbb{R})$ by x and y .

(a) Show that for the natural \mathfrak{g} -valued form ω on G we have

$$\omega = \frac{1}{x} \begin{pmatrix} dx & dy \\ 0 & 0 \end{pmatrix},$$

so that dx/x and dy/x are left invariant 1-forms on G , and a left invariant 2-form is $(dx \wedge dy)/x^2$.

(b) Find the structure constants for these forms.

(c) Show that

$$(dP) \cdot P^{-1} = \frac{1}{x} \begin{pmatrix} dx & -y dx + x dy \\ 0 & 0 \end{pmatrix}$$

and find the right invariant 2-forms.

24. (a) Show that the natural $\mathfrak{gl}(n, \mathbb{R})$ -valued 1-form ω on $\text{GL}(n, \mathbb{R})$ is given by

$$\omega^{ij} = \frac{1}{\det(x^{\alpha\beta})} \sum_{k=1}^n y^{ik} dx^{kj},$$

where

$$(y^{\alpha\beta}) = \det(x^{\alpha\beta}) \cdot (x^{\alpha\beta})^{-1}.$$

(b) Show that both the left and right invariant n^2 -forms are multiples of

$$\frac{1}{(\det(x^{\alpha\beta}))^n} (dx^{11} \wedge \cdots \wedge dx^{n1}) \wedge \cdots \wedge (dx^{1n} \wedge \cdots \wedge dx^{nn}).$$

25. The special linear group $\text{SL}(n, \mathbb{R}) \subset \text{GL}(n, \mathbb{R})$ is the set of all matrices of determinant 1.

(a) Using Problem 15, show that its Lie algebra $\mathfrak{sl}(n, \mathbb{R})$ consists of all matrices with trace = 0.

(b) For the case of $\mathrm{SL}(2, \mathbb{R})$, show that

$$P^{-1} \cdot dP = \begin{pmatrix} v dx - y du & v dy - y dv \\ -u dx + x du & -u dy + x dv \end{pmatrix},$$

where we use x, y, u, v for $x^{11}, x^{12}, x^{21}, x^{22}$. Check that the trace is 0 by differentiating the equation $xv - yu = 1$.

(c) Show that a left invariant 3-form is

$$v dx \wedge du \wedge dy - y dx \wedge du \wedge dv.$$

26. For $M, N \in \mathfrak{o}(n) = \mathcal{L}(\mathrm{O}(n)) = \{M : M = -M^t\}$, define

$$(N, M) = -\mathrm{trace} M \cdot N^t.$$

(a) $(\ , \)$ is a positive definite inner product on $\mathfrak{o}(n)$.

(b) If $A \in \mathrm{O}(n)$, then

$$(\mathrm{Ad}(\bar{A})\bar{M}, \mathrm{Ad}(A)N) = (M, N).$$

($\mathrm{Ad}(A)$ is defined in Problem 19.)

(c) The left invariant metric on $\mathrm{O}(n)$ with value $(\ , \)$ at $\mathrm{O}(n)_I$ is also right invariant.

27. (a) If G is a compact Lie group, then $\exp : \mathfrak{g} \rightarrow G$ is onto. *Hint:* Use Proposition 21.

(b) Let $A \in \mathrm{SL}(2, \mathbb{R})$. Recall that A satisfies its characteristic polynomial, so $A^2 - (\mathrm{trace} A)A + I = 0$. Conclude that $\mathrm{trace} A^2 \geq -2$.

(c) Show that the following element of $\mathrm{SL}(2, \mathbb{R})$ is not A^2 for any A . Conclude that it is not in the image of \exp .

$$\begin{pmatrix} -2 & 0 \\ 0 & -1/2 \end{pmatrix}$$

(d) $\mathrm{SL}(2, \mathbb{R})$ does not have a bi-invariant metric.

28. Let x be a coordinate system around e in a Lie group G , let $\pi_j : G \times G \rightarrow G$ be the projections, and let (y, z) be the coordinate system around (e, e) given by $y^i = x^i \circ \pi_1$, $z^i = x^i \circ \pi_2$. Define $\phi^i : G \times G \rightarrow \mathbb{R}$ by

$$\phi^i(a, b) = x^i(ab),$$

and let X_i be the left invariant vector field on G with

$$X_i(e) = \left. \frac{\partial}{\partial x^i} \right|_e.$$

(a) Show that

$$X_i = \sum_{j=1}^n \psi_i^j \frac{\partial}{\partial x^j},$$

where

$$\psi_i^j(a) = \frac{\partial \phi^j}{\partial x^i}(a, e).$$

(b) Using $L_a L_b = L_{ab}$, show that

$$[L_{a*} X_i(b)](x^l) = [X_i(ab)](x^l).$$

Deduce that

$$X_i(b)(x^l \circ L_a) = \psi_i^l(ab),$$

and then that

$$\sum_{j=1}^n \psi_i^j(b) \cdot \frac{\partial \phi^l}{\partial x^j}(a, b) = \psi_i^l(ab).$$

Letting $\tilde{\psi} = (\tilde{\psi}_j^i)$ be the inverse matrix of $\psi = (\psi_i^j)$, we can write

$$\frac{\partial \phi^l}{\partial x^j}(a, b) = \sum_{i=1}^n \psi_i^l(ab) \cdot \tilde{\psi}_j^i(b).$$

This equation (or any of numerous things equivalent to it) is known as *Lie's first fundamental theorem*. The associativity of G is implicitly contained in it, since we used the fact that $L_a L_b = L_{ab}$.

(c) Prove the *converse of Lie's first fundamental theorem*, which states the following. Let $\phi = (\phi^1, \dots, \phi^n)$ be a differentiable function in a neighborhood of $0 \in \mathbb{R}^{2n}$ [with standard coordinate system $y^1, \dots, y^n, z^1, \dots, z^n$] such that

$$\phi(a, 0) = a \quad \text{for } a \in \mathbb{R}^n.$$

Suppose there are differentiable functions ψ_j^i in a neighborhood of $0 \in \mathbb{R}^n$ [with standard coordinate system x^1, \dots, x^n] such that

$$\psi_j^i(0) = \delta_j^i$$

$$(*) \quad \frac{\partial \phi^l}{\partial z^j}(a, b) = \sum_{i=1}^n \psi_i^l(\phi(a, b)) \cdot \tilde{\psi}_j^i(b) \quad \text{for } (a, b) \text{ in a neighborhood of } 0 \in \mathbb{R}^{2n}.$$

Then $(a, b) \mapsto \phi(a, b)$ is a **local Lie group structure** on a neighborhood of $0 \in \mathbb{R}^n$ (it is associative and has inverses for points close enough to 0, which serves as the identity); the corresponding left invariant vector fields are

$$X_i = \sum_{j=1}^n \psi_i^j \frac{\partial}{\partial x^j}.$$

[To prove associativity, note that

$$\frac{\partial \phi^l(\phi(a, b), z)}{\partial z^j} = \sum_{i=1}^n \psi_i^l(\phi(\phi(a, b), z)) \cdot \tilde{\psi}_j^i(z) \quad \text{by } (*).$$

and then show that $\phi(a, \phi(b, z))$ satisfies the same equation.]

29. *Lie's second fundamental theorem* states that the left invariant vector fields X_i of a Lie group G satisfy

$$[X_i, X_j] = \sum_{k=1}^n C_{ij}^k X_k$$

for certain *constants* C_{ij}^k —in other words, the bracket of two left invariant vector fields is left invariant. The aim of this problem is to prove the *converse of Lie's second fundamental theorem*, which states the following: A Lie algebra of vector fields on a neighborhood of $0 \in \mathbb{R}^n$, which is of dimension n over \mathbb{R} and contains a basis for \mathbb{R}^n_0 , is the set of left invariant vector fields for some local Lie group structure on a neighborhood of $0 \in \mathbb{R}^n$.

(a) Choose X_1, \dots, X_n in the Lie algebra so that $X_i(0) = \partial/\partial x^i|_0$ and set

$$X_i = \sum_{j=1}^n \psi_i^j \frac{\partial}{\partial x^j}.$$

If

$$\omega^i = \sum_{j=1}^n \tilde{\psi}_j^i dx^j,$$

then the ω^i are the dual forms, and consequently

$$d\omega^k = - \sum_{i < j} C_{ij}^k \omega^i \wedge \omega^j \quad C_{ij}^k \text{ constants.}$$

(b) Let $\pi_j: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the projections. Then

$$\pi_2^* \omega^j - \pi_1^* \omega^j = \sum_{l=1}^n (\tilde{\psi}_l^j \circ \pi_2) \left[d(x^l \circ \pi_2) - \sum_{i=1}^n (\psi_i^l \circ \pi_2) \cdot \pi_1^* \omega^i \right].$$

Consequently, the ideal generated by the forms $d(x^l \circ \pi_2) - \sum_{i=1}^n (\psi_i^l \circ \pi_2) \cdot \pi_1^* \omega^i$ is the same as the ideal \mathcal{L} generated by the forms $\pi_2^* \omega^j - \pi_1^* \omega^j$. Using the fact that the C_{jk}^i are constants, show that $d(\mathcal{L}) \subset \mathcal{L}$. Hence $\mathbb{R}^n \times \mathbb{R}^n$ is foliated by n -dimensional manifolds on which the forms $d(x^l \circ \pi_2) - \sum_{i=1}^n (\psi_i^l \circ \pi_2) \cdot \pi_1^* \omega^i$ all vanish.

(c) Conclude, as in the proof of Theorem 17, that for fixed a , there is a function $\Phi_a: \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfying $\Phi_a(0) = a$ and

$$d\Phi_a^l(b) = \sum_{i=1}^n \psi_i^l(\Phi_a(b)) \cdot \omega^i(b),$$

or equivalently,

$$\frac{\partial \Phi_a^l}{\partial x^j}(b) = \sum_{i=1}^n \psi_i^l(\Phi_a(b)) \cdot \tilde{\psi}_j^i(b).$$

Now set $\phi(a, b) = \Phi_a(b)$, and use the converse of Lie's first fundamental theorem.

30. *Lie's third fundamental theorem* states that the C_{jk}^i satisfy equations (1) and (2) on page 396, i.e., that the left invariant vector fields form a Lie algebra under $[\cdot, \cdot]$. The aim of this problem is to prove the *converse of Lie's third fundamental theorem*, which states that any n -dimensional Lie algebra is the Lie algebra for some local Lie group in a neighborhood of $0 \in \mathbb{R}^n$.

Let C_{ij}^k be constants satisfying equations (1) and (2) on page 396. We would like to find vector fields X_1, \dots, X_n on a neighborhood of $0 \in \mathbb{R}^n$ such that $[X_i, X_j] = \sum_{k=1}^n C_{ij}^k X_k$. Equivalently, we want to find forms ω^i with

$$d\omega^k = - \sum_{i < j} C_{ij}^k \omega^i \wedge \omega^j.$$

Then the result will follow from the converse of Lie's second fundamental theorem.

(a) Let h_r^k be functions on $\mathbb{R} \times \mathbb{R}^n$ such that

$$\frac{\partial h_r^k}{\partial t} = \delta_r^k - \sum_{i,j} C_{ij}^k x^i h_r^j$$

$$h_r^k(0, x) = 0.$$

These are equations "depending on the parameters x " (see Problem 5-5(b)). Note that $h_r^k(t, 0) = \delta_r^k t$, so that $h_r^k(1, 0) = \delta_r^k$. Let σ^k be the 1-form on $\mathbb{R} \times \mathbb{R}^n$ defined by

$$\sigma^k = \sum_r h_r^k dx^r.$$

and write

$$d\sigma^k = \lambda^k + (dt \wedge \alpha^k).$$

where λ^k and α^k do not involve dt . Show that

$$\begin{aligned}\lambda^k &= \sum_{i < j} \left(\frac{\partial h_j^k}{\partial x^i} - \frac{\partial h_i^k}{\partial x^j} \right) dx^i \wedge dx^j \\ \alpha^k &= dx^k - \sum_{i,j} C_{ij}^k x^i \sigma^j.\end{aligned}$$

(b) Show that

$$d\lambda^k = dt \wedge \left(- \sum_{i,j} C_{ij}^k dx^i \wedge \sigma^j - \sum_{i,j} C_{ij}^k x^k \lambda^j \right).$$

(c) Let

$$\theta^k = \lambda^k + \frac{1}{2} \sum_{i,j} C_{ij}^k \sigma^i \wedge \sigma^j.$$

Show that

$$\begin{aligned}d\theta^k &= dt \wedge \left(- \sum_{i,j} C_{ij}^k x^i \lambda^j - \sum_{i,j} \sum_{r,s} C_{ij}^k C_{rs}^i x^r \sigma^s \wedge \sigma^j \right) \\ &\quad + \text{terms not involving } dt.\end{aligned}$$

Using

$$\begin{aligned}\sum_{i,j} \sum_{r,s} C_{ij}^k C_{rs}^i \sigma^s \wedge \sigma^j &= \frac{1}{2} \sum_{i,j} \sum_{r,s} (C_{ij}^k C_{rs}^i - C_{is}^k C_{rj}^i) \sigma^s \wedge \sigma^j \\ &= \frac{1}{2} \sum_{i,j} \sum_{r,s} (C_{ij}^k C_{rs}^i + C_{is}^k C_{jr}^i) \sigma^s \wedge \sigma^j,\end{aligned}$$

and equation (2) on page 396, show that

$$\begin{aligned}d\theta^k &= dt \wedge \left(- \sum_{i,j} C_{ij}^k x^i \lambda^j + \frac{1}{2} \sum_{i,j} \sum_{r,s} C_{ir}^k C_{sj}^i x^r \sigma^s \wedge \sigma^j \right) \\ &\quad + \text{terms not involving } dt.\end{aligned}$$

Finally deduce that

$$d\theta^k = dt \wedge - \sum_{j,l} C_{jl}^k x^j \theta^l + \text{terms not involving } dt.$$

(d) We can write

$$\theta^k = \sum_{i < j} g_{ij}^k dx^i \wedge dx^j.$$

where $g_{ij}^k(0, x) = 0$ (Why?). Using (c), show that

$$\frac{\partial g_{ij}^k}{\partial t} = - \sum_{r,s} C_{rs}^k x^r g_{ij}^s.$$

Conclude that $\theta^k = 0$.

(e) We now have

$$\lambda^k = - \frac{1}{2} \sum_{i,j} C_{ij}^k \sigma^i \wedge \sigma^j.$$

$$d\sigma^k = - \frac{1}{2} \sum_{i,j} C_{ij}^k \sigma^i \wedge \sigma^j + (dt \wedge \alpha^k).$$

Show that the forms $\omega^k(x) = \sigma^k(1, x)$ satisfy

$$d\omega^k = - \frac{1}{2} \sum_{i,j} C_{ij}^k \omega^i \wedge \omega^j.$$

CHAPTER 11

EXCURSION IN THE REALM OF ALGEBRAIC TOPOLOGY

This chapter explores further properties of the de Rham cohomology vector spaces of a manifold. Our main results will be restatements, in terms of the de Rham cohomology, of fundamental properties of the ordinary cohomology which is studied in algebraic topology. Because we deal only with manifolds, many of the proofs become significantly easier. On the other hand, we will be using some of the main tools of algebraic topology, thus retaining much of the flavor of that subject. Along the way we will deduce all sorts of interesting consequences, including a theorem about the possibility of imbedding n -manifolds in \mathbb{R}^{n+1} .

Let M be a manifold with $M = U \cup V$ for open sets $U, V \subset M$. Before examining the cohomology of M we will simply look at the vector space $C^k(M)$ of k -forms on M . Let

$$\begin{array}{ll} i_U: U \rightarrow M & i_V: V \rightarrow M \\ j_U: U \cap V \rightarrow U & j_V: U \cap V \rightarrow V \end{array}$$

be the inclusions. Then we have two linear maps α and β .

$$C^k(M) \xrightarrow{\alpha = i_U^* \oplus i_V^*} C^k(U) \oplus C^k(V) \xrightarrow{\beta = j_U^* - j_V^*} C^k(U \cap V)$$

defined by

$$\alpha(\omega) = (i_U^*(\omega), i_V^*(\omega)) \quad \beta(\lambda_1, \lambda_2) = j_U^*(\lambda_1) - j_V^*(\lambda_2).$$

Here $i_U^*(\omega)$ is just the restriction of ω to U , etc. Clearly $\beta \circ \alpha = 0$. In other words, $\text{image } \alpha \subset \ker \beta$. Moreover, the converse holds: $\ker \beta \subset \text{image } \alpha$. For, if $\beta(\lambda_1, \lambda_2) = 0$, then $\lambda_1 = \lambda_2$ on $U \cap V$, so we can define ω on M to be λ_1 on U and λ_2 on V , and then $\alpha(\omega) = (\lambda_1, \lambda_2)$. The equation $\text{image } \alpha = \ker \beta$ is expressed by saying that the above diagram is exact at the middle vector space. We can extend this diagram by putting the vector space containing only 0 at the

ends; the arrows at either end of the following sequence are the only possible linear maps.

1. LEMMA. The sequence

$$0 \rightarrow C^k(M) \xrightarrow{\alpha} C^k(U) \oplus C^k(V) \xrightarrow{\beta} C^k(U \cap V) \rightarrow 0$$

is exact at all places.

PROOF. It is clear that α is one-one. This is equivalent to exactness at $C^k(M)$, since the image of the first map is $\{0\} \subset C^k(M)$. Similarly, exactness at $C^k(U \cap V)$ is equivalent to β being onto. To prove that β is onto, let $\{\phi_U, \phi_V\}$ be a partition of unity subordinate to $\{U, V\}$. Then $\omega \in C^k(U \cap V)$ is

$$\omega = \beta(\phi_V \omega, -\phi_U \omega),$$

where $\phi_V \omega$ denotes the form equal to $\phi_V \omega$ on $U \cap V$, and equal to 0 on $U - (U \cap V)$. ♦

By putting in the maps d , we can expand our diagram as follows.

$$\begin{array}{ccccccc} & \vdots & & \vdots & & \vdots & \\ & \downarrow & & \downarrow & & \downarrow & \\ 0 & \longrightarrow & C^k(M) & \xrightarrow{\alpha} & C^k(U) \oplus C^k(V) & \xrightarrow{\beta} & C^k(U \cap V) \longrightarrow 0 \\ & & \downarrow d & & \downarrow d \oplus d & & \downarrow d \\ 0 & \longrightarrow & C^{k+1}(M) & \xrightarrow{\alpha} & C^{k+1}(U) \oplus C^{k+1}(V) & \xrightarrow{\beta} & C^{k+1}(U \cap V) \longrightarrow 0 \\ & & \downarrow & & \downarrow & & \downarrow \end{array}$$

so that the rows are all exact. It is easy to check that this diagram commutes, that is, any two compositions from one vector space to another are equal:

$$\begin{array}{lcl} (d \oplus d) \circ \alpha = \alpha \circ d & \xrightarrow{\alpha} \downarrow d \oplus d = d \downarrow \xrightarrow{\alpha} & \\ d \circ \beta = \beta \circ (d \oplus d) & \xrightarrow{\beta} \downarrow d = d \oplus d \downarrow \xrightarrow{\beta} & \end{array}$$

Our first main theorem depends only on the simple algebraic structure inherent in this diagram. To isolate this purely algebraic structure, we make the following definitions. A complex C is a sequence of vector spaces C^k , $k = 0, 1, 2, \dots$, together with a sequence of linear maps

$$d^k: C^k \rightarrow C^{k+1}$$

satisfying $d^{k+1} \circ d^k = 0$, or briefly, $d^2 = 0$. A map $\alpha: C_1 \rightarrow C_2$ between complexes is a sequence of linear maps

$$\alpha^k: C_1^k \rightarrow C_2^k$$

such that the following diagram commutes for all k .

$$\begin{array}{ccc} C_1^k & \xrightarrow{\alpha^k} & C_2^k \\ d_1^k \downarrow & & \downarrow d_2^k \\ C_1^{k+1} & \xrightarrow{\alpha^{k+1}} & C_2^{k+1} \end{array}$$

The most important examples of complexes are obtained by choosing $C^k = C^k(M)$ for some manifold M , with d^k the operator d on k -forms. Another example, implicit in our discussion, is the direct sum $C = C_1 \oplus C_2$ of two complexes, defined by

$$C^k = C_1^k \oplus C_2^k, \quad d^k = d_1^k \oplus d_2^k.$$

For any complex C we can define the cohomology vector spaces of C by

$$H^k(C) = \frac{\ker d^k}{\text{image } d^{k-1}}.$$

Naturally, if $C = \{C^k(M)\}$, then $H^k(C)$ is just $H^k(M)$. If $\alpha: C_1 \rightarrow C_2$ is a map between complexes, then we have a map, also denoted by α ,

$$\alpha: H^k(C_1) \rightarrow H^k(C_2).$$

To define α we note that every element of $H^k(C_1)$ is determined by some $x \in C_1^k$ with $d_1^k(x) = 0$. Commutativity of the above diagram shows that $d_2^k(\alpha^k(x)) = \alpha^{k+1}d_1^k(x) = 0$, so $\alpha^k(x)$ determines an element of $H^k(C_2)$, which we define to be α (the class determined by x). This map is well-defined.

for if we change x to $x + d_1^{k-1}(y)$ for some $y \in C_1^{k-1}$, then $\alpha^k(x)$ is changed to

$$\begin{aligned}\alpha^k(x + d_1^{k-1}(y)) &= \alpha^k(x) + \alpha^k(d_1^{k-1}(y)) \\ &= \alpha^k(x) + d_2^{k-1}(\alpha^{k-1}(y)).\end{aligned}$$

which determines the same element of $H^k(C_2)$. When $C_1^k = C^k(M)$, $C_2^k = C^k(N)$, and $\alpha: C^k(M) \rightarrow C^k(N)$ is f^* for $f: N \rightarrow M$, then this map is just $f^*: H^k(M) \rightarrow H^k(N)$.

Now suppose that we have an exact sequence of complexes

$$0 \rightarrow C_1 \xrightarrow{\alpha} C_2 \xrightarrow{\beta} C_3 \rightarrow 0.$$

which really means a vast commutative diagram in which all rows are exact.

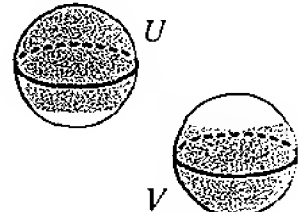
$$(*) \quad \begin{array}{ccccccc} & & \downarrow & & \downarrow & & \downarrow \\ 0 & \longrightarrow & C_1^{k-1} & \xrightarrow{\alpha^{k-1}} & C_2^{k-1} & \xrightarrow{\beta^{k-1}} & C_3^{k-1} \longrightarrow 0 \\ & & \downarrow d_1^{k-1} & & \downarrow d_2^{k-1} & & \downarrow d_3^{k-1} \\ 0 & \longrightarrow & C_1^k & \xrightarrow{\alpha^k} & C_2^k & \xrightarrow{\beta^k} & C_3^k \longrightarrow 0 \\ & & \downarrow d_1^k & & \downarrow d_2^k & & \downarrow d_3^k \\ 0 & \longrightarrow & C_1^{k+1} & \xrightarrow{\alpha^{k+1}} & C_2^{k+1} & \xrightarrow{\beta^{k+1}} & C_3^{k+1} \longrightarrow 0 \\ & & \downarrow & & \downarrow & & \downarrow \end{array}$$

What does this imply about the maps $\alpha: H^k(C_1) \rightarrow H^k(C_2)$ and $\beta: H^k(C_2) \rightarrow H^k(C_3)$? The nicest thing that could happen would be for the following diagram to be exact:

$$0 \rightarrow H^k(C_1) \xrightarrow{\alpha} H^k(C_2) \xrightarrow{\beta} H^k(C_3) \rightarrow 0.$$

This is *not* true. For example, if U and V are overlapping portions of S^2 for which there is a deformation retraction of $U \cap V$ into S^1 , then we have an exact sequence

$$0 \rightarrow C^k(S^2) \rightarrow C^k(U) \oplus C^k(V) \rightarrow C^k(U \cap V) \rightarrow 0.$$



but *not* an exact sequence

$$\begin{array}{ccccccc} 0 & \longrightarrow & H^1(S^2) & \longrightarrow & H^1(U) \oplus H^1(V) & \longrightarrow & H^1(U \cap V) \longrightarrow 0. \\ & & \cong & & \cong & & \cong \\ & & 0 & & 0 & & \mathbb{R} \end{array}$$

Nevertheless, something very nice is true:

2. THEOREM. If $0 \rightarrow C_1 \xrightarrow{\alpha} C_2 \xrightarrow{\beta} C_3 \rightarrow 0$ is a short exact sequence of complexes, then there are linear maps

$$\delta^k : H^k(C_3) \rightarrow H^{k+1}(C_1)$$

so that the following infinitely long sequence is exact (everywhere):

$$\begin{aligned} 0 \longrightarrow H^0(C_1) &\xrightarrow{\alpha} H^0(C_2) \xrightarrow{\beta} H^0(C_3) \xrightarrow{\delta} H^1(C_1) \longrightarrow \dots \\ \dots \longrightarrow H^k(C_1) &\xrightarrow{\alpha} H^k(C_2) \xrightarrow{\beta} H^k(C_3) \xrightarrow{\delta} H^{k+1}(C_1) \longrightarrow \dots \end{aligned}$$

PROOF. Throughout the proof, diagram (*) should be kept at hand. Let $x \in C_3^k$ with $d_3^k(x) = 0$. By exactness of the middle row of (*), there is $y \in C_2^k$ with $\beta^k(y) = x$. Then

$$0 = d_3^k(x) = d_3^k \beta^k(y) = \beta^{k+1} d_2^k(y).$$

So $d_2^k(y) \in \ker \beta^{k+1} = \text{image } \alpha^{k+1}$; thus $d_2^k(y) = \alpha^{k+1}(z)$ for some (unique) $z \in C_1^{k+1}$. Moreover,

$$\alpha^{k+1} d_1^{k+1}(z) = d_2^{k+1} \alpha^{k+1}(z) = d_2^{k+1} d_2^k(y) = 0.$$

Since α^{k+1} is one-one, this implies that $d_1^{k+1}(z) = 0$, so z determines an element of $H^{k+1}(C_1)$; this element is defined to be δ^k of the element of $H^k(C_3)$ determined by x .

In order to prove that δ^k is well-defined, we must check that the result does not depend on the choice of $x \in C_3^k$ representing the element of $H^k(C_3)$. So we have to show that we obtain $0 \in H^{k+1}(C_1)$ if we start with an element of the form $d_3^{k-1}(x')$ for $x' \in C_3^{k-1}$. In this case, let $x' = \beta^{k-1}(y')$. Then

$$x = d_3^{k-1}(x') = d_3^{k-1} \beta^{k-1}(y') = \beta^k d_2^{k-1}(y'),$$

so we choose $d_2^{k-1}(y')$ as y . This means that $d_2^k(y) = 0$, and hence $z = 0$.

It is also necessary to check that our definition is independent of the choice of y with $\beta^k(y) = x$; this is left to the reader.

The proof that the sequence is exact consists of 6 similar diagram chases. We will supply the proof that $\ker \alpha \subset \text{image } \delta$. Let $x \in C_1^k$ satisfy $d_1^k(x) = 0$, and suppose that $\alpha^k(x) \in C_2^k$ represents $0 \in H^k(C_2)$. This means that $\alpha^k(x) = d_2^{k-1}(y)$ for some $y \in C_2^{k-1}$. Now

$$d_3^{k-1} \beta^{k-1}(y) = \beta^k d_2^{k-1}(y) = \beta^k \alpha^k(x) = 0.$$

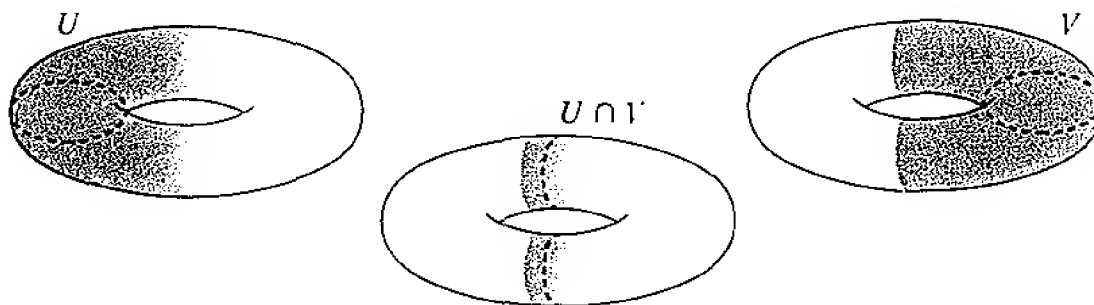
So $\beta^{k-1}(y)$ represents an element of $H^{k-1}(C_3)$. Moreover, the definition of δ immediately shows that the image of this element under δ is precisely the class represented by x . ♦

It is a worthwhile exercise to check that the main step in the proof of Theorem 8-16 is precisely the proof that $\ker \alpha \subset \text{image } \delta$, together with the first part of the proof that δ is well-defined. All of Theorem 8-16 can be derived directly from the following corollary of Lemma 1 and Theorem 2.

3. THEOREM (THE MAYER-VIETORIS SEQUENCE). If $M = U \cup V$, where U and V are open, then we have an exact sequence (eventually ending in 0's):

$$0 \rightarrow H^0(M) \rightarrow \cdots \rightarrow H^k(M) \rightarrow H^k(U) \oplus H^k(V) \rightarrow H^k(U \cap V) \xrightarrow{\delta} H^{k+1}(M) \rightarrow \cdots$$

As several of the Problems show, the cohomology of nearly everything can be computed by a suitable application of the Mayer-Vietoris sequence. As a simple example, we consider the torus $T = S^1 \times S^1$, and the open sets U and V illustrated below. Since there is a deformation retraction of U and V

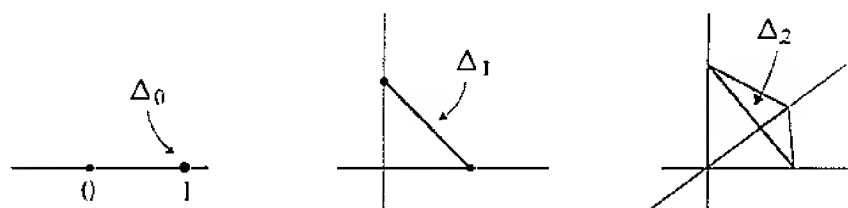


onto circles, and a deformation retraction of $U \cap V$ onto 2 circles, the Mayer-Vietoris sequence is

Rather than compute the cohomology of other manifolds, we will use the Mayer-Vietoris sequence to relate the dimensions of $H^k(M)$ to an entirely different set of numbers, arising from a “triangulation” of M , a new structure which we will now define.

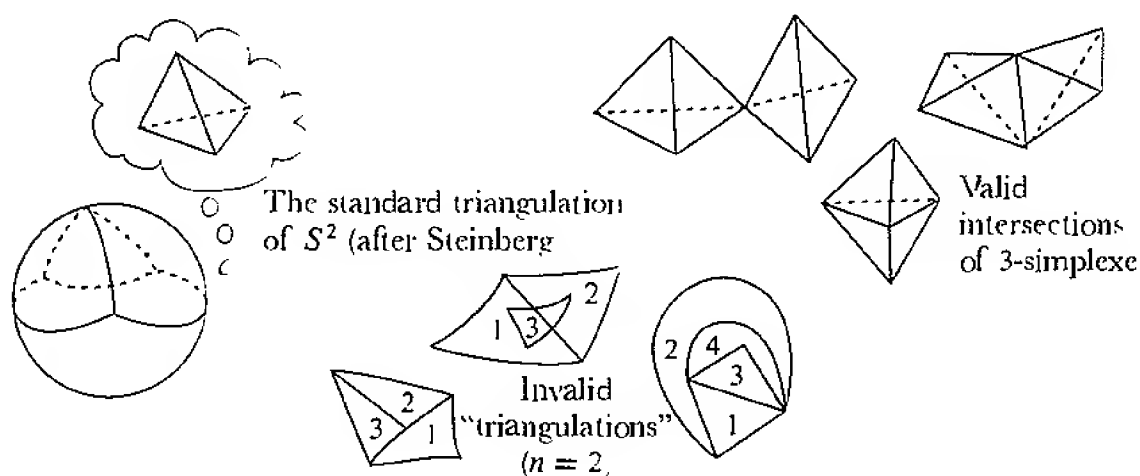
The standard n -simplex Δ_n is defined as the set

$$\Delta_n = \left\{ x \in \mathbb{R}^{n+1} : 0 \leq x^i \leq 1 \text{ and } \sum_{i=1}^{n+1} x^i = 1 \right\}.$$



(In Problem 8-5, Δ_n is defined to be a different, although homeomorphic, set.) The subset of Δ_n obtained by setting $n - k$ of the coordinates x^i equal to 0 is homeomorphic to Δ_k , and is called a k -face of Δ_n . If $\Delta \subset M$ is a diffeomorphic image of some Δ_m , then the image of a k -face of Δ_m is called a k -face of Δ . Now by a triangulation of a compact n -manifold M we mean a finite collection $\{\sigma^n_i\}$ of diffeomorphic images of Δ_n which cover M and which satisfy the following condition:

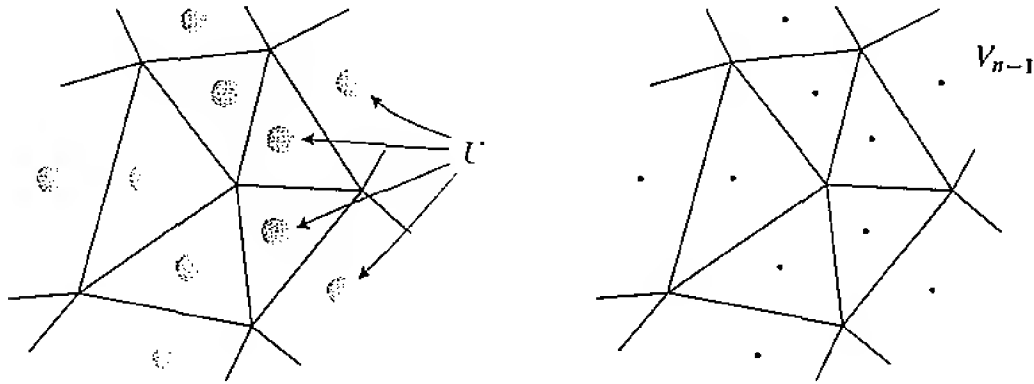
If $\sigma^n_i \cap \sigma^n_j \neq \emptyset$, then for some k the intersection $\sigma^n_i \cap \sigma^n_j$ is a k -face of both σ^n_i and σ^n_j .



It is a difficult theorem that every C^∞ manifold has a triangulation; for a proof see Munkres, *Elementary Differential Topology*, or Whitney, *Geometric Integration*

Theory. Assuming that our manifold M has a triangulation $\{\sigma^n_i\}$ we will call each σ^n_i an n -simplex of the triangulation; any k -face of any σ^n_i will be called a k -simplex of the triangulation. We let α_k be the number of these k -simplexes.

Now let U be the disjoint union of open balls, one within each n -simplex σ^n_i , and let V_{n-1} be the complement of the set consisting of the centers of these balls, so that V_{n-1} is a neighborhood of the union of all $(n-1)$ -simplexes of M . Then



$M = U \cup V_{n-1}$ where $U \cap V_{n-1}$ has the same cohomology as a disjoint union of α_n copies of S^{n-1} . Consider first the case where $n > 2$. The Mayer-Vietoris sequence breaks into pieces:

$$\begin{array}{ccccccc}
 (1) & 0 \longrightarrow & H^0(M) & \longrightarrow & H^0(U) \oplus H^0(V_{n-1}) & \longrightarrow & H^0(U \cap V_{n-1}) \longrightarrow H^1(M) \\
 & & & & \parallel & & \parallel \\
 & & & & 0 & & 0
 \end{array}$$

(2) For $1 < k < n-1$,

$$\begin{array}{ccccccc}
 H^{k-1}(U \cap V_{n-1}) & \longrightarrow & H^k(M) & \longrightarrow & H^k(U) \oplus H^k(V_{n-1}) & \longrightarrow & H^k(U \cap V_{n-1}) \\
 \parallel & & \parallel & & \parallel & & \parallel \\
 0 & & 0 & & 0 & & 0
 \end{array}$$

$$\begin{array}{ccccccc}
 (3) & H^{n-2}(U \cap V_{n-1}) & \longrightarrow & H^{n-1}(M) & \longrightarrow & H^{n-1}(U) \oplus H^{n-1}(V_{n-1}) & \longrightarrow \\
 & \parallel & & \parallel & & \parallel & \\
 & 0 & & 0 & & 0 & \\
 & & & & & & \\
 & \longrightarrow & H^{n-1}(U \cap V_{n-1}) & \longrightarrow & H^n(M) & \longrightarrow & H^n(U) \oplus H^n(V_{n-1}) \\
 & & & & & & \parallel \\
 & & & & & & 0
 \end{array}$$

Applying Proposition 4 to these pieces yields

$$\dim H^k(V_{n-1}) = \dim H^k(M) \quad 0 \leq k \leq n-2$$

$$\dim H^{n-1}(V_{n-1}) = \dim H^{n-1}(M) - \dim H^n(M) + \alpha_n.$$

For the case $n = 2$ we easily obtain the same result without splitting up the sequence. We now introduce the Euler characteristic $\chi(M)$ of M , defined by

$$\chi(M) = \dim H^0(M) - \dim H^1(M) + \dim H^2(M) - \cdots + (-1)^n \dim H^n(M).$$

This makes sense for any manifold in which all $H^k(M)$ are finite dimensional; we anticipate here a later result that $H^k(M)$ is finite dimensional whenever M is compact. The above equations then imply that

$$\begin{aligned} \chi(V_{n-1}) &= \sum_{k=0}^{n-1} (-1)^k \dim H^k(V_{n-1}) \\ &= \sum_{k=0}^{n-2} (-1)^k \dim H^k(M) \\ &\quad + (-1)^{n-1} [\dim H^{n-1}(M) - \dim H^n(M) + \alpha_n] \\ &= \chi(M) - (-1)^n \alpha_n. \end{aligned}$$

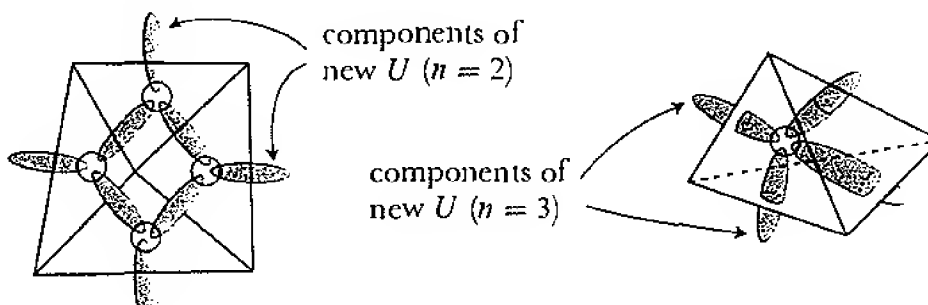
or

$$\chi(M) = \chi(V_{n-1}) + (-1)^n \alpha_n.$$

5. THEOREM. For any triangulation of a compact manifold M we have

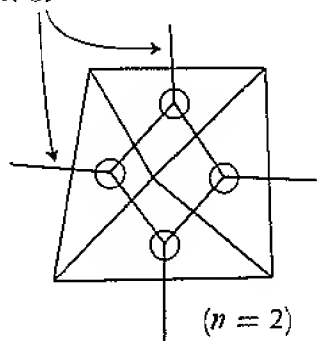
$$\chi(M) = \alpha_0 - \alpha_1 + \alpha_2 - \cdots + (-1)^n \alpha_n.$$

PROOF. In the manifold V_{n-1} we define a new open set U which consists of a disjoint union of sets diffeomorphic to \mathbb{R}^n , one for each $(n-1)$ -face, joining the balls of the old U .

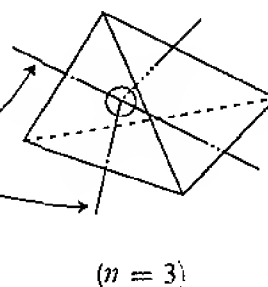


We will let V_{n-2} be the complement of arcs, in the new U , joining the centers of the balls in the old U .

V_{n-2} is the
complement of



V_{n-2} is the
complement of



An argument precisely like that which proves the equation

$$\chi(M) = \chi(V_{n-1}) + (-1)^n \alpha_n$$

also shows that

$$\chi(V_{n-1}) = \chi(V_{n-2}) + (-1)^{n-1} \alpha_{n-1}.$$

Similarly, we introduce V_{n-3}, \dots, V_0 ; the last of these is a disjoint union of α_0 sets each of which is smoothly contractible to a point. Hence $\chi(V_0) = \alpha_0$, while in all other cases we have

$$\chi(V_k) = \chi(V_{k-1}) + (-1)^k \alpha_k.$$

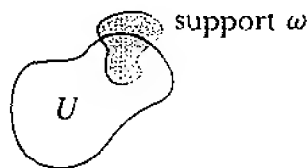
Combining these equations, we have

$$\begin{aligned} \chi(M) &= \chi(V_{n-1}) + (-1)^n \alpha_n \\ &= \chi(V_{n-2}) + [(-1)^{n-1} \alpha_{n-1} + (-1)^n \alpha_n] \\ &\quad \vdots \\ &= \chi(V_0) + [(-1)^1 \alpha_1 + \dots + (-1)^n \alpha_n] \\ &= \alpha_0 - \alpha_1 + \dots + (-1)^n \alpha_n. \quad \spadesuit \end{aligned}$$

6. COROLLARY (DESCARTES-EULER). If a convex polyhedron has V vertices, E edges, and F faces, then

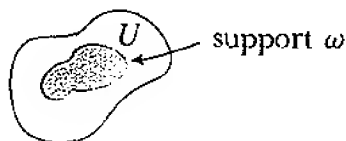
$$V - E + F = 2.$$

If we turn from H^k to H_c^k we encounter a very different situation. If $U \subset M$ is open, a form ω with compact support $\subset M$ may not restrict to a form with compact support $\subset U$: the inclusion map of U into M is not proper. On the



other hand, if ω is a form with compact support $\subset U$, then ω can be extended to M by letting it be 0 outside U ; we will denote this extended form by

$i_{U'}(\omega).$



If $C_c^k(M)$ denotes the vector space of k -forms with compact support on M , we can define a new sequence.

7. LEMMA. The sequence

$$0 \rightarrow C_c^k(U \cap V) \xrightarrow{j_U' \oplus -j_V'} C_c^k(U) \oplus C_c^k(V) \xrightarrow{i_U' + i_V'} C_c^k(M) \rightarrow 0$$

is exact.

PROOF. It is clear that $j_U' \oplus -j_V'$ is one-one; in fact, each map j_U' and j_V' is one-one.

To prove that $i_U' + i_V'$ is onto, let ω be a k -form with compact support on M , and let $\{\phi_U, \phi_V\}$ be a partition of unity for the cover $\{U, V\}$. Then

$$\omega = \phi_U \omega + \phi_V \omega$$

is clearly the image of $(\phi_U \omega, \phi_V \omega) \in C_c^k(U) \oplus C_c^k(V)$.

It is clear that $\text{image}(j_U' \oplus -j_V') \subset \ker(i_U' + i_V')$. To prove the converse, suppose that

$$(\lambda_1, \lambda_2) \in C_c^k(U) \oplus C_c^k(V) \quad \text{satisfies} \quad i_U'(\lambda_1) + i_V'(\lambda_2) = 0.$$

This means that $\lambda_1 = -\lambda_2$. Since $\text{support } \lambda_1 \subset U$ and $\text{support } \lambda_2 \subset U$, this shows that $\text{support } \lambda_1 \subset U \cap V$ and $\text{support } \lambda_2 \subset U \cap V$. So (λ_1, λ_2) is the image of $\lambda_1 \in C_c^k(U \cap V)$. ♦

8. THEOREM (MAYER-VIETORIS FOR COMPACT SUPPORTS). If the manifold $M = U \cup V$ for U, V open in M , then there is a long exact sequence

$$\cdots \rightarrow H_c^k(U \cap V) \rightarrow H_c^k(U) \oplus H_c^k(V) \rightarrow H_c^k(M) \xrightarrow{\delta} H_c^{k+1}(U \cap V) \rightarrow \cdots$$

PROOF. Apply Theorem 2 to the short exact sequence of complexes given by the Lemma. ♦

This sequence is much harder to work with than the Mayer-Vietoris sequence. For example, suppose we want to find H_c^k for $\mathbb{R}^n - \{0\}$, which is diffeomorphic to $S^{n-1} \times \mathbb{R}$. If we write $S^n = U \cup V$ in the usual way, so that $U \cap V$ is diffeomorphic to $S^{n-1} \times \mathbb{R}$, then $S^n \times \mathbb{R} = (U \times \mathbb{R}) \cup (V \times \mathbb{R})$, where $(U \times \mathbb{R}) \cap (V \times \mathbb{R})$ is diffeomorphic to $S^{n-1} \times \mathbb{R}^2$. The only way to use induction is to find H_c^k for all $S^n \times \mathbb{R}^m$, starting with $S^1 \times \mathbb{R}^m$. The details will be left to the reader; we will merely record one further result, for later use, and then proceed to yet another application of Theorem 2.

9. COROLLARY. If $M = U \cup V$ for U, V open in M , then there is a dual long exact sequence

$$\cdots \rightarrow H_c^{k+1}(U \cap V)^* \rightarrow H_c^k(M)^* \rightarrow [H_c^k(U) \oplus H_c^k(V)]^* \rightarrow H_c^k(U \cap V)^* \rightarrow \cdots$$

PROOF. We just have to show that if the sequence of linear maps

$$W_1 \xrightarrow{\alpha} W_2 \xrightarrow{\beta} W_3$$

is exact at W_2 , then so is the sequence of dual maps and spaces

$$W_3^* \xrightarrow{\beta^*} W_2^* \xrightarrow{\alpha^*} W_1^*.$$

For any $\lambda \in W_3^*$ we have

$$\alpha^* \beta^*(\lambda) = \alpha^*(\lambda \circ \beta) = \lambda \circ (\beta \circ \alpha) = \lambda \circ 0 = 0.$$

So $\alpha^* \circ \beta^* = 0$.

Now suppose $\lambda \in W_2^*$ satisfies $\alpha^*(\lambda) = 0$. Then $\lambda \circ \alpha = 0$. We claim that

$$\begin{array}{ccccc} W_1 & \xrightarrow{\alpha} & W_2 & \xrightarrow{\beta} & W_3 \\ & & \downarrow \lambda & \nearrow \bar{\lambda} & \\ & & \mathbb{R} & & \end{array}$$

there is $\bar{\lambda}: W_3 \rightarrow \mathbb{R}$ with $\lambda = \beta^*(\bar{\lambda})$, i.e., $\lambda = \beta \circ \bar{\lambda}$. Given a $w \in W_3$ which is

of the form $\beta(w')$, we define

$$\bar{\lambda}(w) = \lambda(\beta').$$

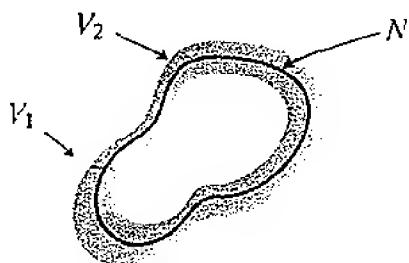
This makes sense, for if $\beta(w') = \beta(w'')$, then $w - w'' = \alpha(z)$ for some z , so $\lambda(w) - \lambda(w'') = \lambda\alpha(z) = 0$. This defines $\bar{\lambda}$ on $\beta(W_2) \subset W_3$. Now choose $W \subset W_3$ with $W_3 = \beta(W_2) \oplus W$, and define $\bar{\lambda}$ to be 0 on W . ♦

We now consider a rather different situation. Let $N \subset M$ be a compact submanifold of M . Then $M - N$ is also a manifold. We therefore have the sequence

$$C_c^k(M - N) \xrightarrow{e} C_c^k(M) \xrightarrow{i^*} C^k(N),$$

where e is “extension”. This sequence is *not* exact at $C_c^k(M)$: the kernel of i^* contains all $\omega \in C_c^k(M)$ which are 0 on N , while the image of e contains all $\omega \in C_c^k(M)$ which are 0 in a neighborhood of N .

To circumvent this difficulty, we will have to use a technical device. We appeal first to a result from the Addendum to Chapter 9. There is a compact neighborhood V of N and a map $\pi: V \rightarrow N$ such that V is a manifold-with-boundary, and if $j: N \rightarrow V$ is the inclusion, then $\pi \circ j$ is the identity of N , while $j \circ \pi$ is smoothly homotopic to the identity of V . We now construct a sequence of such neighborhoods $V = V_1 \supset V_2 \supset V_3 \supset \dots$ with $\bigcap_i V_i = N$.



Now consider two forms $\omega_i \in C^k(V_i)$, $\omega_j \in C^k(V_j)$. We will call ω_i and ω_j *equivalent* if there is $l > i, j$ such that

$$\omega_i|_{V_l} = \omega_j|_{V_l}.$$

It is clear that we can make the set of all equivalence classes into a vector space $\mathcal{G}^k(N)$, the “germs of k -forms in a neighborhood of M ”. Moreover, it is easy to define $d: \mathcal{G}^k(N) \rightarrow \mathcal{G}^{k+1}(N)$, so that we obtain a complex \mathcal{G} . Finally, we define a map of complexes

$$C_c^k(M) \xrightarrow{i^*} \mathcal{G}^k(N)$$

in the obvious way: $\omega \mapsto$ the equivalence class of any $\omega|_{V_i}$.

10. LEMMA. The sequence

$$0 \rightarrow C_c^k(M - N) \xrightarrow{e} C_c^k(M) \xrightarrow{i^*} \mathcal{G}^k(N) \rightarrow 0$$

is exact.

PROOF. Clearly e is one-one.

If $\omega \in C_c^k(M - N)$, then $\omega = 0$ in some neighborhood U of N . Since N is compact and $\bigcap_i V_i = N$, there is some i such that $V_i \subset U$, and consequently $\omega = 0$ on V_i . This means that $i^*e(\omega) = 0$. Conversely, suppose $\lambda \in C_c^k(M)$ satisfies $i^*(\lambda) = 0$. By definition of $\mathcal{G}^k(N)$, this means that $\lambda|_{V_i} = 0$ for some i . Hence $\lambda|_{M - N}$ has compact support $\subset M - N$, and $\lambda = e(\lambda|_{M - N})$.

Finally, any element of $\mathcal{G}^k(N)$ is represented by a form η on some V_i . Let $f: M \rightarrow [0, 1]$ be a C^∞ function which is 1 on V_{i+1} , having support $f \subset \text{interior } V_i$. Then $f\eta \in C_c^k(M)$, and $f\eta$ represents the same element of $\mathcal{G}^k(N)$ as η ; consequently this element is $i^*(f\eta)$. ♦

11. LEMMA. The cohomology vector spaces $H^k(\mathcal{G})$ of the complex $\{\mathcal{G}^k(N)\}$ are isomorphic to $H^k(N)$ for all k .

PROOF. This follows easily from the fact that $j^*: H^k(V_i) \rightarrow H^k(N)$ is an isomorphism for each V_i . Details are left to the reader. ♦

12. THEOREM (THE EXACT SEQUENCE OF A PAIR). If $N \subset M$ is a compact submanifold of M , then there is an exact sequence

$$\dots \rightarrow H_c^k(M - N) \rightarrow H_c^k(M) \rightarrow H^k(N) \xrightarrow{\delta} H_c^{k+1}(M - N) \rightarrow \dots$$

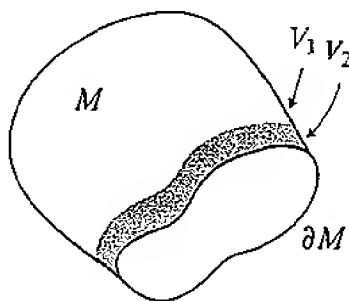
PROOF. Apply Theorem 2 to the exact sequence of complexes given by Lemma 10, and then use Lemma 11. ♦

In the proof of this theorem, the de Rham cohomology of the manifold-with-boundary V_i entered only as an intermediary (and we could have replaced the V_i by their interiors). But in the next theorem, which we will need later, it is the object of primary interest.

13. THEOREM. Let M be a manifold-with-boundary, with compact boundary ∂M . Then there is an exact sequence

$$\dots \rightarrow H_c^k(M - \partial M) \rightarrow H_c^k(M) \rightarrow H^k(\partial M) \xrightarrow{\delta} H_c^{k+1}(M - \partial M) \rightarrow \dots$$

PROOF. Just like the proof of Theorem 12, using tubular neighborhoods V_i of ∂M in M . ♦



As a simple application of Theorem 13, we can rederive $H_c^k(\mathbb{R}^n)$ from a knowledge of $H^k(S^{n-1})$, by choosing M to be the closed ball B in \mathbb{R}^n , with $H_c^k(B) \approx H^k(B) = 0$ for $k \neq 0$. The reader may use Theorem 12 to compute $H_c^k(S^n \times \mathbb{R}^m)$, by considering the pair $(S^n \times \mathbb{R}^m, \{p\} \times \mathbb{R}^m)$. Then Theorem 13 may be used to compute the cohomology of $S^n \times S^{m-1} = \partial(S^n \times \text{closed ball in } \mathbb{R}^m)$. For our next application we will seek bigger game.

Let $M \subset \mathbb{R}^{n+1}$ be a compact n -dimensional submanifold of \mathbb{R}^{n+1} (a compact "hypersurface" of \mathbb{R}^{n+1}). Using Theorem 8-17, the sequence of the pair (\mathbb{R}^{n+1}, M) gives

$$\begin{array}{ccccccc} H_c^n(\mathbb{R}^{n+1}) & \longrightarrow & H^n(M) & \xrightarrow{\delta} & H_c^{n+1}(\mathbb{R}^{n+1} - M) & \longrightarrow & H_c^{n+1}(\mathbb{R}^{n+1}) \longrightarrow H^{n+1}(M). \\ \parallel & & & & \cong & & \parallel \\ 0 & & & & \mathbb{R} & & 0 \end{array}$$

It follows that

$$(*) \quad \text{number of components of } \mathbb{R}^{n+1} - M = \dim H^n(M) + 1.$$

But we also know (Problem 8-25) that

$$(**) \quad \text{number of components of } \mathbb{R}^{n+1} - M \geq 2.$$

14. THEOREM. If $M \subset \mathbb{R}^{n+1}$ is a compact hypersurface, then M is orientable, and $\mathbb{R}^{n+1} - M$ has exactly 2 components. Moreover, M is the boundary of each component.

PROOF. From (*) and (**) we obtain

$$\dim H^n(M) + 1 \geq 2.$$

Since $\dim H^n(M)$ is either 0 or 1, we conclude that $\dim H^n(M) = 1$, so M is orientable; then (*) shows that $\mathbb{R}^{n+1} - M$ has exactly two components. The proof in Problem 8-25 shows that every point of M is arbitrarily close to points in different components of $\mathbb{R}^{n+1} - M$, so every point of M is in the boundary of each of the two components. ♦

15. COROLLARY (GENERALIZED $[C^\infty]$ JORDAN CURVE THEOREM). If $M \subset \mathbb{R}^{n+1}$ is a submanifold homeomorphic to S^n , then $\mathbb{R}^{n+1} - M$ has two components, and M is the boundary of each.

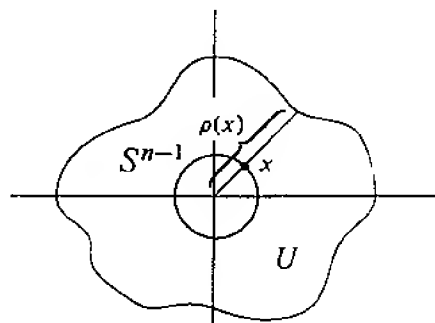
16. COROLLARY. Neither the projective plane nor the Klein bottle can be imbedded in \mathbb{R}^3 .

Our next main result will combine some of the theorems we already have. However, there are a number of technicalities involved, which we will have to dispose of first.

Consider a bounded open set $U \subset \mathbb{R}^n$ which is star-shaped with respect to 0. Then U can be described as

$$U = \{tx : x \in S^{n-1} \text{ and } 0 \leq t < \rho(x)\}$$

for a certain function $\rho: S^{n-1} \rightarrow \mathbb{R}$. We will call ρ the radial function of U .

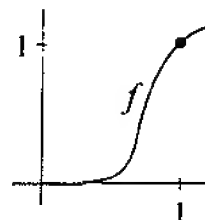


If ρ is C^∞ , then we can prove that U is diffeomorphic to the open ball B of radius 1 in \mathbb{R}^n . The basic idea of the proof is to take $tx \in B$ to $\rho(x)t \cdot x \in U$. This produces difficulties at 0, so a modification is necessary.

17. LEMMA. If the radial function ρ of a star-shaped open set $U \subset \mathbb{R}^n$ is C^∞ , then U is diffeomorphic to the open ball B of radius 1 in \mathbb{R}^n .

PROOF. We can assume, without loss of generality, that $\rho \geq 1$ on S^{n-1} . Let $f: [0, 1] \rightarrow [0, 1]$ be a C^∞ function with

$$\begin{aligned} f &= 0 \text{ in a neighborhood of } 0 \\ f' &\geq 0 \\ f(1) &= 1. \end{aligned}$$

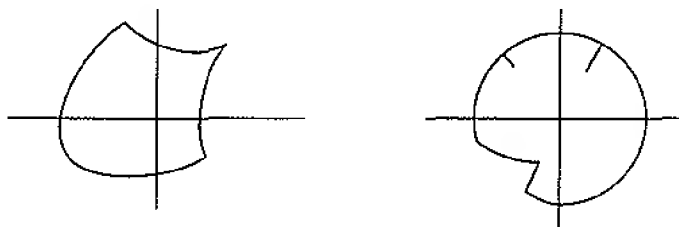


Define $h: B \rightarrow U$ by

$$h(tx) = [t + (\rho(x) - 1)f(t)]x, \quad x \in S^{n-1}, \quad 0 \leq t < 1.$$

Clearly h is a one-one map of B onto U . It is the identity in a neighborhood of 0, so it is C^∞ , with a non-zero Jacobian, at 0. At any other point the same conclusion follows from the fact that $t \mapsto t + (\rho(x) - 1)f(t)$ is a C^∞ function with strictly positive derivative. ♦

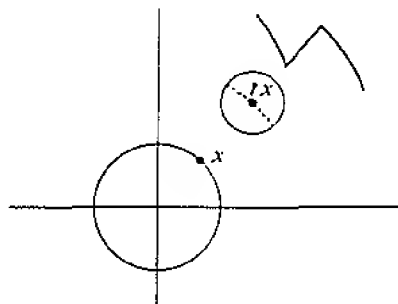
In general, the function ρ need not be C^∞ ; it might not even be continuous.



However, the discontinuities of ρ can be of a certain form only.

18. LEMMA. At each point $x \in S^{n-1}$, the radial function ρ of a star-shaped open set $U \subset \mathbb{R}^n$ is “lower semi-continuous”: for every $\varepsilon > 0$ there is a neighborhood W of x in S^{n-1} such that $\rho(y) > \rho(x) - \varepsilon$ for all $y \in W$.

PROOF. Choose $tx \in U$ with $\rho(x) - t < \varepsilon$. Since U is open, there is an open



ball B with $x \in B \subset U$. There is clearly a neighborhood W of x with the property that for $y \in W$ the point ty is in B , and hence in U . This means that for $y \in W$ we have $\rho(y) \geq t > \rho(x) - \varepsilon$. ♦

Even when ρ is discontinuous, it looks as if U should be diffeomorphic to \mathbb{R}^n . Proving this turns out to be quite a feat, and we will be content with proving the following.

19. LEMMA. If U is an open star-shaped set in \mathbb{R}^n , then $H^k(U) \approx H^k(\mathbb{R}^n)$ and $H_c^k(U) \approx H_c^k(\mathbb{R}^n)$ for all k .

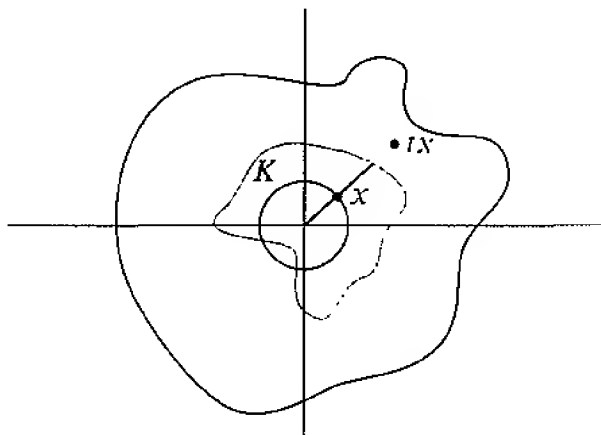
PROOF. The proof for H^k is clear, since U is smoothly contractible to a point. We also know that $H_c^n(U) \approx \mathbb{R} \approx H_c^n(\mathbb{R}^n)$. By Theorem 8-17, we just have to show that $H_c^k(U) = 0$ for $0 \leq k < n$.

Let ω be a closed k -form with compact support $K \subset U$. We claim that there is a C^∞ function $\bar{\rho}: S^{n-1} \rightarrow \mathbb{R}$ such that $\bar{\rho} < \rho$ and

$$K \subset V = \{tx : x \in S^{n-1} \text{ and } 0 \leq t < \bar{\rho}(x)\}.$$

This will prove the Lemma, for then V is diffeomorphic to \mathbb{R}^n , and consequently $\omega = d\eta$ where η has compact support contained in V , and hence in U .

For each $x \in S^{n-1}$, choose $t_x < \rho(x)$ such that all points in K of the form ux for $0 \leq u \leq \rho(x)$ actually have $u < t_x$. Since K is closed and ρ is lower semi-



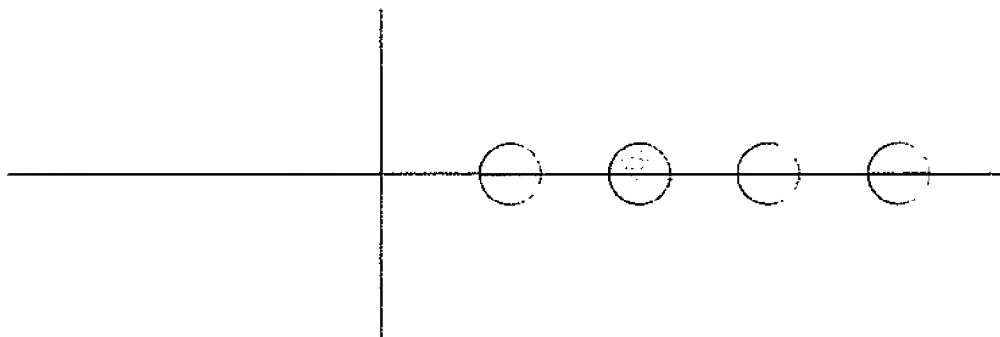
continuous, there is a neighborhood W_x of x in S^{n-1} such that t_x may also be used as t_y for all $y \in W$. Let W_{x_1}, \dots, W_{x_r} cover S^{n-1} , let ϕ_1, \dots, ϕ_r be a partition of unity subordinate to this cover, and define

$$\bar{\rho} = t_{x_1}\phi_1 + \dots + t_{x_r}\phi_r.$$

Any point $x \in S^{n-1}$ is in a certain subcollection of the W_{x_i} , say W_{x_1}, \dots, W_{x_l} for convenience. Then $\rho_{l+1}(x), \dots, \rho_r(x)$ are 0. Each t_{x_1}, \dots, t_{x_l} is $< \rho(x)$. Since $\phi_1(x) + \dots + \phi_l(x) = 1$, it follows that $\bar{\rho}(x) < \rho(x)$. Similarly, $K \subset V$. ♦

We can apply this last Lemma in the following way. Let M be a compact manifold, and choose a Riemannian metric for M . According to Problem 9-32, every point has a neighborhood U which is geodesically convex; we can also choose U so that for any $p \in U$ the map \exp_p takes an open subset of M_p diffeomorphically onto U . Let $\{U_1, \dots, U_r\}$ be a finite cover by such open sets. If any $V = U_{i_1} \cap \dots \cap U_{i_l}$ is non-empty, then V is clearly geodesically convex. If $p \in V$, then \exp_p establishes a diffeomorphism of V with an open star-shaped set in M_p . It follows from Lemma 19 that V has the same H^k and H_c^k as \mathbb{R}^n . In general, a manifold M will be called of **finite type** if there is a finite cover $\{U_1, \dots, U_r\}$ such that each non-empty intersection has the same H^k and H_c^k as \mathbb{R}^n ; such a cover will be called **nice**.

It is fairly clear that if we consider $\mathbb{N} = \{1, 2, 3, \dots\}$ as a subset of \mathbb{R}^2 , then $M = \mathbb{R}^2 - \mathbb{N}$ is not of finite type. To prove this rigorously, we first use the Mayer-Vietoris sequence for $\mathbb{R}^2 = M \cup V$, where V is a disjoint union of balls around $1, 2, 3, \dots$. We obtain



$$\begin{array}{ccccccc}
 H^1(\mathbb{R}^2) & \longrightarrow & H^1(M) \oplus H^1(V) & \longrightarrow & H^1(M \cap V) & \longrightarrow & H^2(\mathbb{R}^2), \\
 \parallel & & \parallel & & & & \parallel \\
 0 & & 0 & & & & 0
 \end{array}$$

where $M \cap V$ has the same H^1 as a disjoint union of infinitely many copies of S^1 ; this shows that $H^1(M)$ is infinite dimensional (see Problem 7 for more information about the cohomology of M). On the other hand,

20. PROPOSITION. If M has finite type, then $H^k(M)$ and $H_c^k(M)$ are finite dimensional for all k .

PROOF. By induction on the number of open sets r in a nice cover. It is clear for $r = 1$. Suppose it is true for a certain r , and consider a nice cover $\{U_1, \dots, U_r, U\}$ of M . Then the theorem is true for $V = U_1 \cup \dots \cup U_r$ and

for U . It is also true for $U \cap V$, since this has the nice cover $\{U \cap U_1, \dots, U \cap U_r\}$. Now consider the Mayer-Vietoris sequence

$$\dots \rightarrow H^{k-1}(U \cap V) \xrightarrow{\delta} H^k(M) \xrightarrow{\alpha} H^k(U) \oplus H^k(V) \rightarrow \dots$$

The map α maps $H^k(M)$ onto a finite dimensional vector space, and the kernel of α is also finite dimensional. So $H^k(M)$ must be finite dimensional.

The proof for $H_c^k(M)$ is similar. ♦

For any manifold M we can define (see Problem 8-31) the **cup product** map

$$H^k(M) \times H^l(M) \xrightarrow{\cup} H^{k+l}(M)$$

by

$$([\omega], [\eta]) \mapsto [\omega \wedge \eta].$$

We can also define

$$H^k(M) \times H_c^l(M) \xrightarrow{\cup} H_c^{k+l}(M)$$

by the same formula, since $\omega \wedge \eta$ has compact support if η does. Now suppose that M^n is connected and oriented, with orientation μ . There is then a unique element of $H_c^n(M)$ represented by any $\eta \in C_c^n(M)$ with

$$\int_{(M, \mu)} \eta = 1.$$

It is convenient to also use μ to denote both this element of $H_c^n(M)$ and the isomorphism $H_c^n(M) \rightarrow \mathbb{R}$ which takes this element to $1 \in \mathbb{R}$. Now every $\alpha \in H^k(M)$ determines an element of the dual space $H_c^{n-k}(M)^*$ by

$$\beta \mapsto \alpha \cup \beta \in H_c^n(M) \xrightarrow{\mu} \mathbb{R}.$$

We denote this element of $H_c^{n-k}(M)^*$ by $PD(\alpha)$, the “Poincaré dual” of α , so that we have a map

$$PD: H^k(M) \rightarrow H_c^{n-k}(M)^*, \quad PD(\alpha)(\beta) = \mu(\alpha \cup \beta).$$

One of the fundamental theorems of manifold theory states that PD is always an isomorphism. We are all set up to prove this fact, but we shall restrict the theorem to manifolds of finite type, in order not to plague ourselves with additional technical details. As with most big theorems of algebraic topology, the main part of the proof is called a Lemma, and the theorem itself is a simple corollary.

21. LEMMA. If $M = U \cup V$ for open sets U and V and PD is an isomorphism for all k on U , V , and $U \cap V$, then PD is also an isomorphism for all k on M .

PROOF. Let $l = n - k$. Consider the following diagram, in which the top row is the Mayer-Vietoris sequence, and the bottom row is the dual of the Mayer-Vietoris sequence for compact supports.

$$\begin{array}{ccccccccc}
 H^{k-1}(U) \oplus H^{k-1}(V) & \longrightarrow & H^{k-1}(U \cap V) & \longrightarrow & H^k(M) & \longrightarrow & H^k(U) \oplus H^k(V) & \longrightarrow & H^k(U \cap V) \\
 \downarrow PD \oplus PD & & \downarrow PD & & \downarrow PD & & \downarrow PD \oplus PD & & \downarrow PD \\
 [H_c^{l+1}(U) \oplus H_c^{l+1}(V)]^* & \longrightarrow & H_c^{l+1}(U \cap V)^* & \longrightarrow & H_c^l(M)^* & \longrightarrow & [H_c^l(U) \oplus H_c^l(V)]^* & \longrightarrow & H_c^l(U \cap V)^*
 \end{array}$$

By assumption, all vertical maps, except possibly the middle one, are isomorphisms. It is not hard to check (Problem 8) that every square in this diagram commutes up to sign, so that by changing some of the vertical isomorphisms to their negatives, we obtain a commutative diagram. We now forget all about our manifold and use a purely algebraic result.

“THE FIVE LEMMA”. Consider the following commutative diagram of vector spaces and linear maps. Suppose that the rows are exact, and that $\phi_1, \phi_2, \phi_4, \phi_5$ are isomorphisms. Then ϕ_3 is also an isomorphism.

$$\begin{array}{ccccccccc}
 V_1 & \xrightarrow{\alpha_1} & V_2 & \xrightarrow{\alpha_2} & V_3 & \xrightarrow{\alpha_3} & V_4 & \xrightarrow{\alpha_4} & V_5 \\
 \downarrow \phi_1 & & \downarrow \phi_2 & & \downarrow \phi_3 & & \downarrow \phi_4 & & \downarrow \phi_5 \\
 W_1 & \xrightarrow{\beta_1} & W_2 & \xrightarrow{\beta_2} & W_3 & \xrightarrow{\beta_3} & W_4 & \xrightarrow{\beta_4} & W_5
 \end{array}$$

PROOF. Suppose $\phi_3(x) = 0$ for some $x \in V_3$. Then $\beta_3\phi_3(x) = 0$, so $\phi_4\alpha_3(x) = 0$. Hence $\alpha_3(x) = 0$, since ϕ_4 is an isomorphism. By exactness at V_3 , there is $y \in V_2$ with $x = \alpha_2(y)$. Thus $0 = \phi_3(x) = \phi_3\alpha_2(y) = \beta_2\phi_2(y)$. Hence $\phi_2(y) = \beta_1(z)$ for some $z \in W_1$. Moreover, $z = \phi_1(w)$ for some $w \in V_1$. Then

$$\phi_2(y) = \beta_1(z) = \beta_1\phi_1(w) = \phi_2\alpha_1(w),$$

which implies that $y = \alpha_1(w)$. Hence

$$x = \alpha_2(y) = \alpha_2(\alpha_1(w)) = 0.$$

So ϕ_3 is one-one.

The proof that ϕ_3 is onto is similar, and is left to the reader. This proves the original Lemma. ♦

22. THEOREM (THE POINCARÉ DUALITY THEOREM). If M is a connected oriented n -manifold of finite type, then the map

$$PD: H^k(M) \rightarrow H_c^{n-k}(M)^*$$

is an isomorphism for all k .

PROOF. By induction on the number r of open sets in a nice cover of M . The theorem is clearly true for $r = 1$. Suppose it is true for a certain r , and consider a nice cover $\{U_1, \dots, U_r, U\}$ of M . Let $V = U_1 \cup \dots \cup U_r$. The theorem is true for U , V , and for $U \cap V$ (as in the proof of Proposition 19). By the Lemma, it is true for M . This completes the induction step. ♦

23. COROLLARY. If M is a connected oriented n -manifold of finite type, then $H^k(M)$ and $H_c^{n-k}(M)$ have the same dimension.

PROOF. Use the Theorem and Proposition 19, noting that V^* is isomorphic to V if V is finite dimensional. ♦

Even though the Poincaré Duality Theorem holds for manifolds which are not of finite type, Corollary 23 does not. In fact, Problem 7 shows that $H^1(\mathbb{R}^2 - \mathbb{N})$ and $H_c^1(\mathbb{R}^2 - \mathbb{N})$ have different (infinite) dimensions.

24. COROLLARY. If M is a compact connected orientable n -manifold, then $H^k(M)$ and $H^{n-k}(M)$ have the same dimension.

25. COROLLARY. If M is a compact orientable odd-dimensional manifold, then $\chi(M) = 0$.

PROOF. In the expression for $\chi(M)$, the terms $(-1)^k \dim H^k(M)$ and

$$(-1)^{n-k} \dim H^{n-k}(M) = (-1)^{k+1} \dim H^{n-k}(M)$$

cancel in pairs. ♦

A more involved use of Poincaré duality will eventually allow us to say much more about the Euler characteristic of any compact connected oriented manifold M^n . We begin by considering a smooth k -dimensional orientable vector bundle $\xi = \pi: E \rightarrow M$ over M . Orientations μ for M and ν for ξ give an orientation $\mu \oplus \nu$ for the $(n+k)$ -manifold E , since E is locally a product. If $\{U_1, \dots, U_r\}$ is a nice cover of M by geodesically convex sets so small that each bundle $\xi|_{U_i}$ is trivial, then a slight modification of the proof for Lemma 19

shows that $\{\pi^{-1}(U_1), \dots, \pi^{-1}(U_r)\}$ is a nice cover of E , so E is a manifold of finite type. Notice also that for the maps

$$M \begin{array}{c} \xrightarrow{s = 0\text{-section}} \\ \xleftarrow{\pi} \end{array} E$$

we have

$$\pi \circ s = \text{identity of } M$$

$$s \circ \pi \quad \text{is smoothly homotopic to identity of } E,$$

so $\pi^*: H^l(M) \rightarrow H^l(E)$ is an isomorphism for all l . The Poincaré duality theorem shows that there is a unique class $U \in H_c^k(E)$ such that

$$\pi^*\mu \cup U = \mu \oplus \nu \in H_c^{n+k}(E).$$

This class U is called the **Thom class** of ξ . Our first goal will be to find a simpler property to characterize U .

Let $F_p = \pi^{-1}(p)$ be the fibre of ξ over any point $p \in M$, and let $j_p: F_p \rightarrow E$ be the inclusion map. Since j_p is proper, there is an element $j_p^*U \in H_c^k(F_p)$. On the other hand, the orientation ν for ξ determines an orientation ν_p for F_p , and hence an element $\nu_p \in H_c^k(F_p)$.

26. THEOREM. Let (M, μ) be a compact connected oriented manifold, and $\xi = \pi: E \rightarrow M$ an oriented k -plane bundle over M with orientation ν . Then the Thom class U is the unique element of $H_c^k(E)$ with the property that for all $p \in M$ we have $j_p^*U = \nu_p$. (This condition means that

$$\int_{(F_p, \nu_p)} j_p^*\omega = 1,$$

where U is the class of the closed form ω .)

PROOF. Pick some closed form $\omega \in C_c^k(E)$ representing U , and let $\eta \in C^n(M)$ be a form representing μ , so that $\int_{(M, \mu)} \eta = 1$. Our definition of U states that

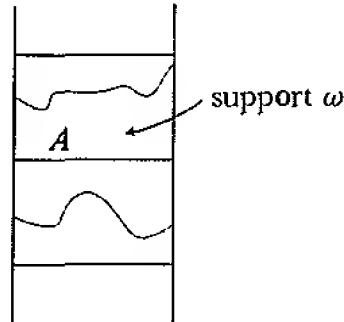
$$(1) \quad \int_E \pi^*\eta \wedge \omega = 1.$$

Let $A \subset M$ be an open set which is diffeomorphic to \mathbb{R}^n , so that A is smoothly contractible to any point $p \in A$. Also choose A so that there is an equivalence

$$f: \pi^{-1}(A) \rightarrow A \times \mathbb{R}^k.$$

This equivalence allows us to identify $\pi^{-1}(A)$ with $A \times F_p$. Under this identification, the map $j_p: F_p \rightarrow \pi^{-1}(A)$ corresponds to the map $e \mapsto (p, e)$ for $e \in F_p$, which we will continue to denote by j_p . We will also use $\pi_2: A \times F_p \rightarrow F_p$ to denote projection on the second factor.

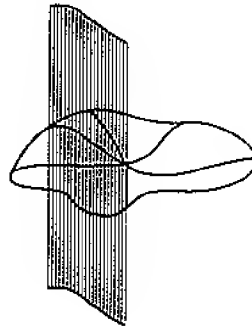
Let $\| \cdot \|$ be a norm on F_p . By choosing a smaller A if necessary, we can assume that there is some $K > 0$ such that, under the identification of $\pi^{-1}(A)$ with $A \times F_p$, the support of $\omega|_{\pi^{-1}(A)}$ is contained in $\{(q, e) : q \in A, \|e\| < K\}$.



Using the fact that A is smoothly contractible to p , it is easy to see that there is a smooth homotopy $H: (A \times F_p) \times [0, 1] \rightarrow A \times F_p$ such that

$$\begin{aligned} H(e, 0) &= e \\ H(e, 1) &= (p, \pi_2(e)) = j_p(\pi_2(e)); \end{aligned}$$

we just pull the fibres along the smooth homotopy which makes A contractible



to e . For the H constructed in this way it follows that

$$H(e, t) \notin \text{support } \omega \text{ if } \|e\| \geq K.$$

Consequently, the form $H^*\omega$ on $(A \times F_p) \times [0, 1]$ has support contained in $\{(q, e, t) : \|e\| < K\}$. A glance at the definition of I (page 224) shows that the

form $IH^*\omega$ on $A \times F_p$ has support contained in $\{(q, e) : \|e\| < K\}$. Theorem 7-14 shows that

$$\begin{aligned}(j_p \circ \pi_2)^*\omega - \omega &= i_1^*(H^*\omega) - i_0^*(H^*\omega) \\ &= d(IH^*\omega) + I(dH^*\omega) \\ &= d(IH^*\omega).\end{aligned}$$

Thus

$$(2) \quad \pi_2^*j_p^*\omega - \omega = d\lambda, \quad \text{support } \lambda \subset \{(q, e) : \|e\| < K\}.$$

So

$$(3) \quad \int_{A \times F_p} \pi^*\eta \wedge \omega = \int_{A \times F_p} \pi^*\eta \wedge \pi_2^*j_p^*\omega - \int_{A \times F_p} \pi^*\eta \wedge d\lambda.$$

Now, on the one hand we have (Problem 8-17)

$$(4) \quad \int_{A \times F_p} \pi^*\eta \wedge \pi_2^*j_p^*\omega = \int_A \pi^*\eta \cdot \int_{F_p} j_p^*\omega.$$

On the other hand, we claim that the last integral in (3) is 0. To prove this, it clearly suffices to prove that the integral is 0 over $A' \times F_p$ for any closed ball $A' \subset A$. Since

$$\pi^*\mu \wedge d\lambda = \pm d(\pi^*\mu \wedge \lambda),$$

we have

$$\begin{aligned}(5) \quad \int_{A' \times F_p} \pi^*\mu \wedge d\lambda &= \pm \int_{A' \times F_p} d(\pi^*\mu \wedge \lambda) && \text{where } \pi^*\mu \wedge \lambda \text{ has} \\ & && \text{compact support on} \\ & && A' \times F_p \text{ by (2)} \\ &= \pm \int_{\partial A' \times F_p} \pi^*\mu \wedge \lambda && \text{by Stokes' Theorem} \\ &= 0.\end{aligned}$$

because the form $\pi^*\mu \wedge \lambda$ is clearly 0 on $\partial A' \times F_p$ (since $\partial A'$ is $(n-1)$ -dimensional).

Combining (3), (4), (5) we see that

$$\int_{A \times F_p} \pi^*\eta \wedge \omega = \int_A \pi^*\eta \cdot \int_{F_p} j_p^*\omega.$$

This shows that $\int_{F_p} j_p^* \omega$ is independent of p , for $p \in A$. Using connectedness, it is easy to see that it is independent of p for all $p \in M$, so we will denote it simply by $\int_F j^* \omega$. Thus

$$\int_{\pi^{-1}(A)} \pi^* \eta \wedge \omega = \int_A \pi^* \eta \cdot \int_F j^* \omega.$$

Comparing with equation (1), and utilizing partitions of unity, we conclude that

$$\int_F j^* \omega = 1,$$

which proves the first part of the theorem.

Now suppose we have another class $U' \in H_c^k(E)$. Since

$$H_c^k(E) \approx H^n(E) \approx H^n(M) \approx \mathbb{R},$$

it follows that $U' = cU$ for some $c \in \mathbb{R}$. Consequently,

$$j_p^* U' = j_p^* cU = c \cdot v_p.$$

Hence U' has the same property as U only if $c = 1$. ♦

The Thom class U of $\xi = \pi: E \rightarrow M$ can now be used to determine an element of $H^k(M)$. Let $s: M \rightarrow E$ be any section; there always is one (namely, the 0-section) and any two are clearly smoothly homotopic. We define the Euler class $\chi(\xi) \in H^k(M)$ of ξ by

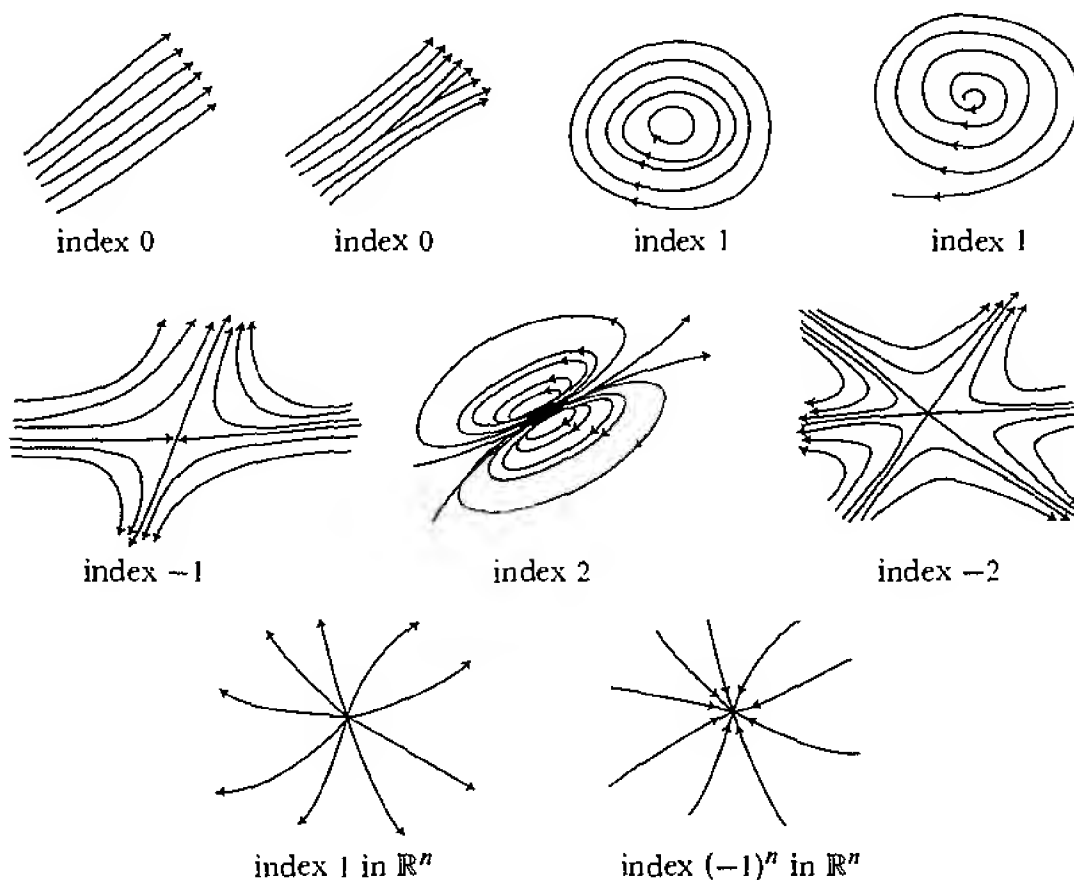
$$\chi(\xi) = s^* U.$$

Notice that if ξ has a non-zero section $s: M \rightarrow E$, and $\omega \in C_c^k(E)$ represents U , then a suitable multiple $c \cdot s$ of s takes M to the complement of support ω . Hence, in this case

$$\chi(\xi) = (c \cdot s)^* U = 0.$$

The terminology “Euler class” is connected with the special case of the bundle TM , whose sections are, of course, vector fields on M . If X is a vector field on M which has an isolated 0 at some point p (that is, $X(p) = 0$, but $X(q) \neq 0$ for $q \neq p$ in a neighborhood of p), then, quite independently of our previous considerations, we can define an “index” of X at p . Consider first a vector

field X on an open set $U \subset \mathbb{R}^n$ with an isolated zero at $0 \in U$. We can define a function $f_X: U - \{0\} \rightarrow S^{n-1}$ by $f_X(p) = X(p)/|X(p)|$. If $i: S^{n-1} \rightarrow U$ is $i(p) = \varepsilon p$, mapping S^{n-1} into U , then the map $f_X \circ i: S^{n-1} \rightarrow S^{n-1}$ has a certain degree; it is independent of ε , for small ε , since the maps $i_1, i_2: S^{n-1} \rightarrow U$ corresponding to ε_1 and ε_2 will be smoothly homotopic. This degree is called the index of X at 0.



Now consider a diffeomorphism $h: U \rightarrow V \subset \mathbb{R}^n$ with $h(0) = 0$. Recall that h_*X is the vector field on V with

$$(h_*X)(y) = h_*(X_{h^{-1}(y)}).$$

Clearly 0 is also an isolated zero of h_*X .

27. LEMMA. If $h: U \rightarrow V \subset \mathbb{R}^n$ is a diffeomorphism with $h(0) = 0$, and X has an isolated 0 at 0, then the index of h_*X at 0 equals the index of X at 0.

PROOF. Suppose first that h is orientation preserving. Define

$$H: \mathbb{R}^n \times [0, 1] \rightarrow \mathbb{R}^n$$

by

$$H(x, t) = \begin{cases} h(tx) & 0 < t \leq 1 \\ Dh(0)(x) & t = 0. \end{cases}$$

This is a smooth homotopy; to prove that it is smooth at 0 we use Lemma 3-2 (compare Problem 3-32). Each map $H_t = x \mapsto H(x, t)$ is clearly a diffeomorphism, $0 \leq t \leq 1$. Note that $H_1 \in \text{SO}(n)$, since h is orientation preserving. There is also a smooth homotopy $\{H_t\}$, $1 \leq t \leq 2$ with each $H_t \in \text{SO}(n)$ and $H_2 = \text{identity}$, since $\text{SO}(n)$ is connected. So (see Problem 8-25), the map h is smoothly homotopic to the identity, via maps which are diffeomorphisms. This shows that f_{h_*X} is smoothly homotopic to f_X on a sufficiently small region of $\mathbb{R}^n - \{0\}$. Hence the degree of $f_{h_*X} \circ i$ is the same as the degree of $f_X \circ i$.

To deal with non-orientation preserving h , it obviously suffices to check the theorem for $h(x) = (x^1, \dots, x^{n-1}, -x^n)$. In this case

$$f_{h_*X} = h \circ f_X \circ h^{-1},$$

which shows that $\text{degree } f_{h_*X} \circ i = \text{degree } f_X \circ i$. ♦

As a consequence of Lemma 27, we can now define the index of a vector field on a manifold. If X is a vector field on a manifold M , with an isolated zero at $p \in M$, we choose a coordinate system (x, U) with $x(p) = 0$, and define the index of X at p to be the index of x_*X at 0.

28. THEOREM. Let M be a compact connected manifold with an orientation μ , which is, by definition, also an orientation for the tangent bundle $\xi = \pi: TM \rightarrow M$. Let $X: M \rightarrow TM$ be a vector field with only a finite number of zeros, and let σ be the sum of the indices of X at these zeros. Then

$$\chi(\xi) = \sigma \cdot \mu \in H^n(M).$$

PROOF. Let p_1, \dots, p_r be the zeros of X . Choose disjoint coordinate systems $(U_1, x_1), \dots, (U_r, x_r)$ with $x_i(p_i) = 0$, and let

$$B_i = x_i^{-1}(\{p \in \mathbb{R}^n : |p| \leq 1\}).$$

If $\omega \in C_c^n(E)$ is a closed form representing the Thom class U of ξ , then we are trying to prove that

$$\int_{(M, \mu)} X^*(\omega) = \sigma.$$

We can clearly suppose that $X(q) \notin \text{support } \omega$ for $q \notin \bigcup_i B_i$. So

$$\int_M X^*(\omega) = \sum_{i=1}^r \int_{B_i} X^*(\omega);$$

thus it suffices to prove that

$$(*) \quad \int_{B_i} X^*(\omega) = \text{index of } X \text{ at } p_i.$$

It will be convenient to drop the subscript i from now on.

We can assume that TM is trivial over B , so that $\pi^{-1}(B)$ can be identified with $B \times M_p$. Let j_p and π_2 have the same meaning as in the proof of Theorem 26. Also choose a norm $\| \cdot \|$ on M_p . We can assume that under the identification of $\pi^{-1}(B)$ with $B \times M_p$, the support of $\omega|_{\pi^{-1}(B)}$ is contained in $\{(q, v) : q \in A, \|v\| \leq 1\}$. Recall from the proof of Theorem 26 that

$$\pi_2^* j_p^* \omega - \omega = d\lambda \quad \text{support } \lambda \subset \{(q, v) : \|v\| \leq 1\}.$$

Since we can assume that $X(q) \notin \text{support } \lambda$ for $q \in \partial B$, we have

$$\begin{aligned} (1) \quad \int_B X^*(\omega) &= \int_B X^* \pi_2^* (j_p^* \omega) - \int_B X^*(d\lambda) \\ &= \int_B X^* \pi_2^* (j_p^* \omega) - \int_{\partial B} X^*(\lambda) \quad \text{by Stokes' Theorem} \\ &= \int_B X^* \pi_2^* (j_p^* \omega). \end{aligned}$$

On the manifold M_p we have

$$j_p^* \omega = d\rho \quad \begin{array}{l} \rho \text{ an } (n-1)\text{-form on } M_p \\ \text{(with non-compact support).} \end{array}$$

If $D \subset M_p$ is the unit disc (with respect to the norm $\| \cdot \|$) and S^{n-1} denotes $\partial D \subset M_p$, then

$$\begin{aligned} (2) \quad \int_{S^{n-1}} \rho &= \int_{\partial D} \rho = \int_D d\rho \\ &= \int_D j_p^* \omega \\ &= 1. \quad \begin{array}{l} \text{by Theorem 26, and the fact} \\ \text{that support } j_p^* \omega \subset D. \end{array} \end{aligned}$$

Now, for $q \in B - \{p\}$, we can define

$$\bar{X}(q) = X(q)/|X(q)|,$$

and $\bar{X}: \partial B \rightarrow TM$ is smoothly homotopic to $X: \partial B \rightarrow TM$. So

$$\begin{aligned} (3) \quad \int_B X^* \pi_2^* (j_p^* \omega) &= \int_B X^* \pi_2^* d\rho \\ &= \int_{\partial B} X^* \pi_2^* \rho \quad \text{by Stokes' Theorem} \\ &= \int_{\partial B} \bar{X}^* \pi_2^* \rho \\ &= \int_{\partial B} (\pi_2 \circ \bar{X})^* \rho. \end{aligned}$$

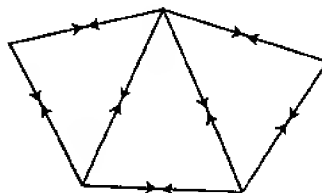
From the definition of the index of a vector field, together with equation (2), it follows that

$$(4) \quad \int_{\partial B} (\pi_2 \circ \bar{X})^* \rho = \text{index of } X \text{ at } p.$$

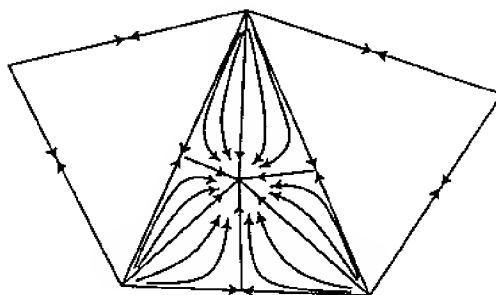
Equations (1), (3), (4) together imply (*). ♦

29. COROLLARY. If X and Y are two vector fields with only finitely many zeros on a compact orientable manifold, then the sum of the indices of X equals the sum of the indices of Y .

At the moment, we do not even know that there is a vector field on M with finitely many zeros, nor do we know what this constant sum of the indices is (although our terminology certainly suggests a good guess). To resolve these questions, we consider once again a triangulation of M . We can then find a vector field X with just one zero in each k -simplex of the triangulation. We begin by drawing the integral curves of X along the 1-simplexes, with a zero at each 0-simplex and at one point in each 1-simplex. We then extend this picture



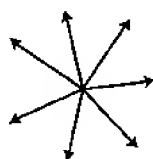
to include the integral curves of X on the 2-simplexes, producing a zero at one



point in each of them. We then continue similarly until the n -simplexes are filled.

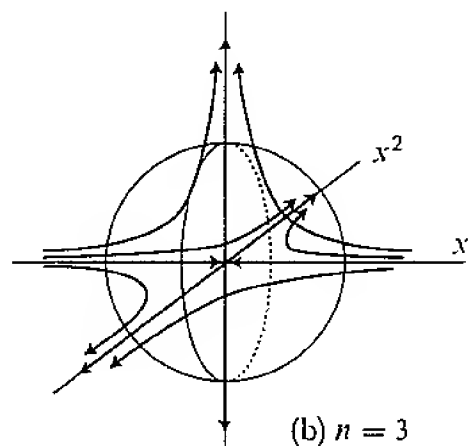
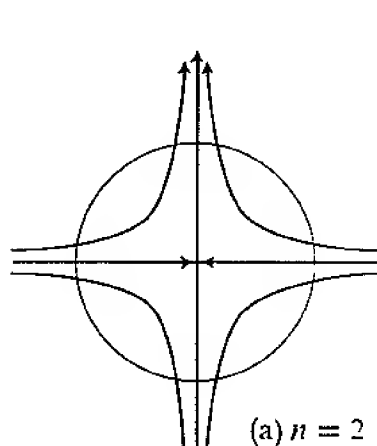
30. THEOREM (POINCARÉ-HOPF). The sum of the indices of this vector field (and hence of any vector field) on M is the Euler characteristic $\chi(M)$. Thus, for $\xi = \pi : TM \rightarrow M$ we have $\chi(\xi) = \chi(M) \cdot \mu$.

PROOF. At each 0-simplex of the triangulation, the vector field looks like

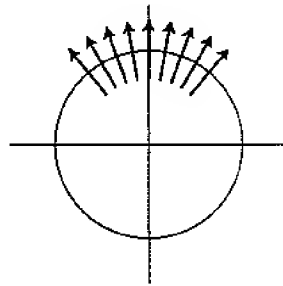


with index 1.

Now consider the vector field in a neighborhood of the place where it is zero on a 1-simplex. The vector field looks like a vector field on $\mathbb{R}^n = \mathbb{R}^1 \times \mathbb{R}^{n-1}$ which points directly inwards on $\mathbb{R}^1 \times \{0\}$ and directly outwards on $\{0\} \times \mathbb{R}^{n-1}$.



For $n = 2$, the index is clearly -1 . To compute the index in general, we note that f_X takes the “north pole” $N = (0, \dots, 0, 1)$ to itself and no other point goes to N . By Theorem 8-12 we just have to compute $\text{sign}_N f_X$. Now at N we can pick projection on $\mathbb{R}^{n-1} \times \{0\}$ as the coordinate system. Along the inverse image of the x^1 -axis the vector field looks exactly like figure (a) above, where we already know the degree is -1 , so f_{X*} takes the subspace of S^{n-1}_N consisting of tangent vectors to this curve into the same subspace, in an orientation reversing way. Along the inverse image of the x^2, \dots, x^{n-1} -axes the vector field looks like



so f_{X*} takes the corresponding subspaces of S^{n-1}_N into themselves in an orientation preserving way. Thus $\text{sign}_N f_X = -1$, which is therefore the index of the vector field.

In general, near a zero within a k -simplex, X looks like a vector field on $\mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k}$ which points directly inwards on $\mathbb{R}^k \times \{0\}$ and directly outwards on $\{0\} \times \mathbb{R}^{n-k}$. The same argument shows that the index is $(-1)^k$.

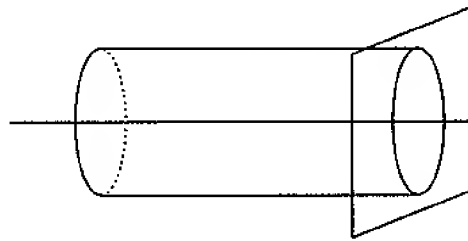
Consequently, the sum of the indices is

$$\alpha_0 - \alpha_1 + \alpha_2 - \dots = \chi(M). \quad \spadesuit$$

We end this chapter with one more observation, which we will need in the last chapter of Volume V! Let $\xi = \pi: E \rightarrow M$ be a smooth oriented k -plane bundle over a compact connected oriented n -manifold M , and let (\cdot, \cdot) be a Riemannian metric for ξ . Then we can form the “associated disc bundle” and “associated sphere bundle”

$$D = \{e : (e, e) \leq 1\}$$

$$S = \{e : (e, e) = 1\}.$$



It is easy to see that D is a compact oriented $(n+k)$ -manifold, with $\partial D = S$; moreover, the D constructed for any other Riemannian metric is diffeomorphic to this one. We let $\pi_0: S \rightarrow M$ be $\pi|_S$.

31. THEOREM. A class $\alpha \in H^k(M)$ satisfies $\pi_0^*(\alpha) = 0$ if and only if α is a multiple of $\chi(\xi)$.

PROOF. Consider the following picture. The top row is the exact sequence

$$\begin{array}{ccccc}
 H_c^k(D - S) & \xrightarrow{e} & H^k(D) & \xrightarrow{i^*} & H^k(S) \\
 & \searrow s^* & \downarrow \bar{s}_* \parallel (\pi|D)^* & \nearrow \pi_0^* & \\
 & & H^k(M) & &
 \end{array}$$

for (D, S) given by Theorem 13. The map $s: M \rightarrow D - S$ is the 0-section, while $\bar{s}: M \rightarrow D$ is the same 0-section. Note that everything commutes.

$$\begin{aligned}
 \pi_0^* &= i^* \circ (\pi|D)^* && \text{since } \pi_0 = (\pi|D) \circ i, \\
 s^* &= \bar{s}_* \circ e && \text{since extending a form to } D \\
 &&& \text{does not affect its value on } s(M),
 \end{aligned}$$

and that

$$\bar{s}^* \circ (\pi|D)^* = \text{identity of } H^k(M),$$

since $(\pi|D) \circ \bar{s}$ is smoothly homotopic to the identity.

Now let $\alpha \in H^k(M)$ satisfy $\pi_0^*(\alpha) = 0$. Then $i^*(\pi|D)^*\alpha = 0$, so $(\pi|D)^*\alpha \in \text{image } e$. Since $D - S$ is diffeomorphic to E , and every element of $H_c^k(D - S)$ is a multiple of the Thom class U of ξ , we conclude that

$$(\pi|D)^*\alpha = c \cdot e(U) \quad \text{for some } c \in \mathbb{R}.$$

Hence

$$\begin{aligned}
 \alpha &= \bar{s}^*(\pi|D)^*\alpha = c \cdot \bar{s}^*(e(U)) = c \cdot s^*U \\
 &= c \cdot \chi(\xi).
 \end{aligned}$$

The proof of the converse is similar. ♦

PROBLEMS

1. Find $H^k(S^1 \times \cdots \times S^1)$ by induction on the number n of factors. [Answer: $\dim H^k = \binom{n}{k}$.]

2. (a) Use the Mayer-Vietoris sequence to determine $H^k(M - \{p\})$ in terms of $H^k(M)$, for a connected manifold M .

(b) If M and N are two connected n -manifolds, let $M \# N$ be obtained by joining M and N as shown below. Find the cohomology of $M \# N$ in terms of that of M and N .



(c) Find χ for the n -holed torus. [Answer: $2 - 2n$.]

3. (a) Find $H^k(\text{Möbius strip})$.

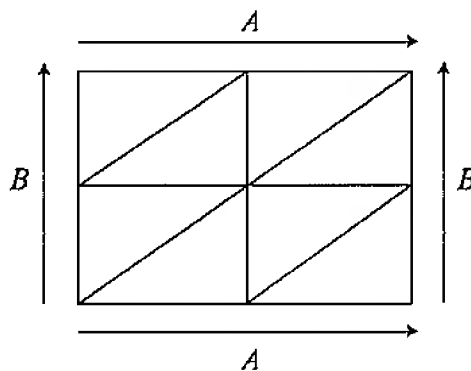
(b) Find $H^k(\mathbb{P}^2)$.

(c) Find $H^k(\mathbb{P}^n)$. (Use Problem 1-15(b); it is necessary to consider whether a neighborhood of \mathbb{P}^{n-1} in \mathbb{P}^n is orientable or not.) [Answer: $\dim H^k(\mathbb{P}^n) = 1$ if k even and $\leq n$, $= 0$ otherwise.]

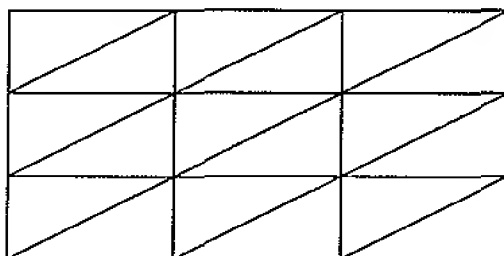
(d) Find $H^k(\text{Klein bottle})$.

(e) Find the cohomology of $M \# (\text{Möbius strip})$ and $M \# (\text{Klein bottle})$ if M is the n -holed torus.

4. (a) The figure below is a triangulation of a rectangle. If we perform the indicated identifications of edges we do *not* obtain a triangulation of the torus. Why not?



(b) The figure below does give a triangulation of the torus when sides are identified. Find α_0 , α_1 , α_2 for this triangulation; compare with Theorem 5 and Problem 1.



5. (a) For any triangulation of a compact 2-manifold M , show that

$$3\alpha_2 = 2\alpha_1$$

$$\alpha_1 = 3(\alpha_0 - \chi(M))$$

$$\frac{\alpha_0(\alpha_0 - 1)}{2} \geq \alpha_1$$

$$\alpha_0 \geq \frac{1}{2}(7 + \sqrt{49 - 24\chi(M)}).$$

(b) Show that for triangulations of S^2 and the torus $T^2 = S^1 \times S^1$ we have

$$S^2 : \quad \alpha_0 \geq 4 \quad \alpha_1 \geq 6 \quad \alpha_2 \geq 4$$

$$T^2 : \quad \alpha_0 \geq 7 \quad \alpha_1 \geq 21 \quad \alpha_2 \geq 14.$$

Find triangulations for which these inequalities are all equalities.

6. (a) Find $H_c^k(S^n \times \mathbb{R}^m)$ by induction on n , using the Mayer-Vietoris sequence for compact supports.

(b) Use the exact sequence of the pair $(S^n \times \mathbb{R}^m, \{p\} \times \mathbb{R}^m)$ to compute the same vector spaces.

(c) Compute $H^k(S^n \times S^{m-1})$, using Theorem 13.

7. (a) The vector space $H^1(\mathbb{R}^2 - \mathbb{N})$ may be described as the set of all sequences of real numbers. Using the exact sequence of the pair $(\mathbb{R}^2, \mathbb{N})$, show that $H_c^1(\mathbb{R}^2 - \mathbb{N})$ may be considered as the set of all real sequences $\{a_n\}$ such that $a_n = 0$ for all but finitely many n .

(b) Describe the map $PD: H^1(\mathbb{R}^2 - \mathbb{N}) \rightarrow H_c^1(\mathbb{R}^2 - \mathbb{N})^*$ in terms of these descriptions of $H^1(\mathbb{R}^2 - \mathbb{N})$ and $H_c^1(\mathbb{R}^2 - \mathbb{N})$, and show that it is an isomorphism.

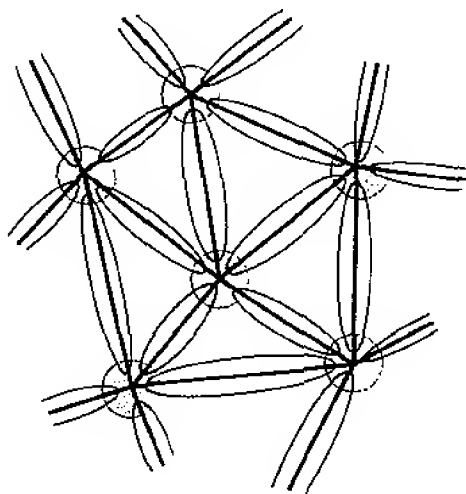
(c) Clearly $H_c^1(\mathbb{R}^2 - \mathbb{N})$ has a countable basis. Show that $H^1(\mathbb{R}^2 - \mathbb{N})$ does not. *Hint:* If $v_i = \{a_i^j\} \in H^1(\mathbb{R}^2 - \mathbb{N})$, choose $(b_1, b_2) \in \mathbb{R}^2$ linearly independent of (a_1^1, a_1^2) ; then choose $(b_3, b_4, b_5) \in \mathbb{R}^3$ linearly independent of both (a_1^3, a_1^4, a_1^5) and (a_2^3, a_2^4, a_2^5) ; etc.



8. Show that the squares in the diagram in the proof of Lemma 21 commute, except for the square

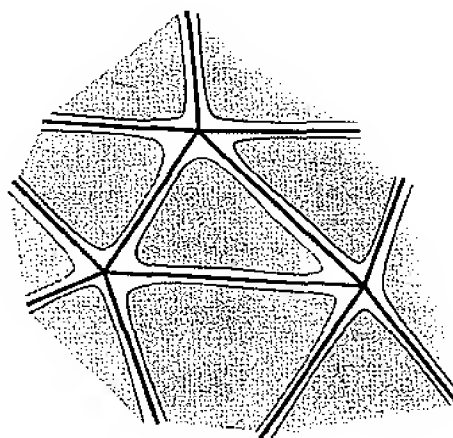
$$\begin{array}{ccc} H^{k-1}(U \cap V) & \longrightarrow & H^k(M) \\ \downarrow PD & & \downarrow PD \\ H_c^{l+1}(U \cap V)^* & \longrightarrow & H_c^l(M)^* \end{array}$$


which commutes up to the sign $(-1)^k$. (It will be necessary to recall how various maps are defined, which is a good exercise; the only slightly difficult maps are the ones involved in the above diagram.)

9. (a) Let $M = M_1 \cup M_2 \cup M_3 \cup \dots$ be a disjoint union of oriented n -manifolds. Show that $H_c^k(M) \approx \bigoplus_i H_c^k(M_i)$, this “direct sum” consisting of all sequences $(\alpha_1, \alpha_2, \alpha_3, \dots)$ with $\alpha_i \in H_c^k(M_i)$ and all but finitely many $\alpha_i = 0 \in H_c^k(M_i)$.
 (b) Show that $H^k(M) \approx \prod_i H^k(M_i)$, this “direct product” consisting of all sequences $(\alpha_1, \alpha_2, \alpha_3, \dots)$ with $\alpha_i \in H^k(M_i)$.
 (c) Show that if the Poincaré duality theorem holds for each M_i , then it holds for M .
 (d) The figure below shows a decomposition of a triangulated 2-manifold into three open sets U_0 , U_1 , and U_2 . Use an analogous decomposition in n dimensions to prove that Poincaré duality holds for any triangulated manifold.



U_0 is union of shaded 
 U_1 is union of unshaded 



U_2 is union of shaded 

10. Let $\xi = \pi: E \rightarrow M$ and $\xi' = \pi': E' \rightarrow M$ be oriented k -plane bundles over a compact oriented manifold M , and (\tilde{f}, f) a bundle map from ξ' to ξ which is an isomorphism on each fibre.

- (a) If $U \in H_c^k(E)$ and $U' \in H_c^k(E')$ are the Thom classes, then $\tilde{f}^*(U) = U'$.
 (b) $f^*(\chi(\xi)) = \chi(\xi')$. (Using the notation of Problem 3-23, we have $f^*(\chi(\xi)) = \chi(f^*(\xi))$.)

11. (a) Let $\xi = \pi: E \rightarrow M$ be an oriented k -plane bundle over an oriented manifold M , with Thom class U . Using Poincaré duality, prove the Thom Isomorphism Theorem: The map $H^l(E) \rightarrow H_c^{l+k}(E)$ given by $\alpha \mapsto \alpha \cup U$ is an isomorphism for all l .

(b) Since we can also consider U as being in $H^k(E)$, we can form $U \cup U \in H_c^{2k}(E)$. Using anticommutativity of \wedge , show that this is 0 for k odd. Conclude that U represents $0 \in H^k(E)$, so that $\chi(\xi) = 0$. It follows, in particular, that $\chi(\xi) = 0$ when $\xi = \pi: TM \rightarrow M$ for M of odd dimension, providing another proof that $\chi(M) = 0$ in this case.

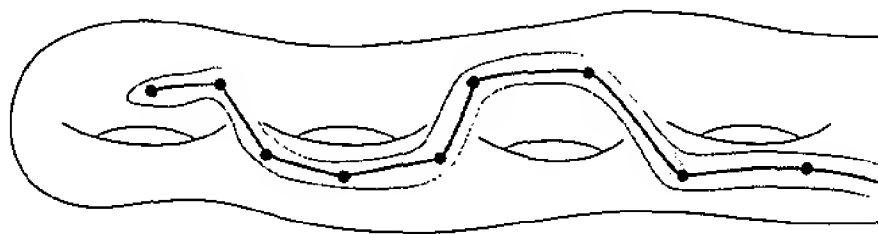
12. If a vector field X has an isolated singularity at $p \in M^n$, show that the index of $-X$ at p is $(-1)^n$ times the index of X at p . This provides another proof that $\chi(M) = 0$ for odd n .

13. (a) Let $p_1, \dots, p_r \in M$. Using Problem 8-26, show that there is a subset $D \subset M$ diffeomorphic to the closed ball, such that all $p_i \in \text{interior } D$.

(b) If M is compact, then there is a vector field X on M with only one singularity.

(c) It is a fact that a C^∞ map $f: S^{n-1} \rightarrow S^{n-1}$ of degree 0 is smoothly homotopic to a constant map. Using this, show that if $\chi(M) = 0$, then there is a nowhere 0 vector field on M .

(d) If M is connected and not compact, then there is a nowhere 0 vector field on M . (Begin with a triangulation to obtain a vector field with a discrete set of zeros. Join these by a ray going to infinity, enclose this ray in a cone, and push everything off to infinity.)



(e) If M is a connected manifold-with-boundary, with $\partial M \neq \emptyset$, then there is a nowhere zero vector field on M .

14. This Problem proves de Rham's Theorem. Basic knowledge of singular cohomology is required. We will denote the group of singular k -chains of X by $S_k(X)$. For a manifold M , we let $S_k^\infty(M)$ denote the C^∞ singular k -chains, and let $i: S_k^\infty(M) \rightarrow S_k(M)$ be the inclusion. It is not hard to show that there is a chain map $\tau: S_k(M) \rightarrow S_k^\infty(M)$ so that $\tau \circ i = \text{identity of } S_k^\infty(M)$, while $i \circ \tau$ is chain homotopic to the identity of $S_k(M)$ [basically, τ is approximation by a C^∞ chain]. This means that we obtain the correct singular cohomology of M if we consider the complex $\text{Hom}(S_k^\infty(M), \mathbb{R})$.

(a) If ω is a closed k -form on M , let $Rh(\omega) \in \text{Hom}(S_k^\infty(M), \mathbb{R})$ be

$$Rh(\omega)(c) = \int_c \omega.$$

Show that Rh is a chain map from $\{C^k(M)\}$ to $\{\text{Hom}(S_k^\infty(M), \mathbb{R})\}$. (*Hint: Stokes' Theorem.*) It follows that there is an induced map Rh from the de Rham cohomology of M to the singular cohomology of M .

(b) Show that Rh is an isomorphism on a smoothly contractible manifold (Lemmas 17, 18, and 19 will not be necessary for this.)

(c) Imitate the proof of Theorem 21, using the Mayer-Vietoris sequence for singular cohomology, to show that if Rh is an isomorphism for U , V , and $U \cap V$, then it is an isomorphism for $U \cup V$.

(d) Conclude that Rh is an isomorphism if M is of finite type. (Using the method of Problem 9, it follows that Rh is an isomorphism for any triangulated manifold.)

(e) Check that the cup product defined using \wedge corresponds to the cup product defined in singular cohomology.

APPENDIX A

CHAPTER 1

Following the suggestions in this chapter, we will now define a manifold to be a topological space M such that

- (1) M is Hausdorff,
- (2) For each $x \in M$ there is a neighborhood U of x and an integer $n \geq 0$ such that U is homeomorphic to \mathbb{R}^n .

Condition (1) is necessary, for there is even a 1-dimensional “manifold” which is not Hausdorff. It consists of $\mathbb{R} \cup \{*\}$ where $* \notin \mathbb{R}$, with the following topology: A set U is open if and only if

- (1) $U \cap \mathbb{R}$ is open,
- (2) If $* \in U$, then $(U \cap \mathbb{R}) \cup \{0\}$ is a neighborhood of 0 (in \mathbb{R}).

Thus the neighborhoods of $*$ look just like neighborhoods of 0. This space may also be obtained by identifying all points except 0 in one copy of \mathbb{R} with the corresponding point in another copy of \mathbb{R} . Although non-Hausdorff manifolds are important in certain cases, we will not consider them.

We have just seen that the Hausdorff property is not a “local property”, but local compactness is, so every manifold is locally compact. Moreover, a Hausdorff locally compact space is regular, so every manifold is regular. (By the way, this argument does not work for “infinite dimensional” manifolds, which are locally like Banach spaces; these need not be regular even if they are Hausdorff.) On the other hand, there are manifolds which are not normal (Problem 6). Every manifold is also clearly locally connected, so every component is open, and thus a manifold itself. Before exhibiting non-metrizable manifolds, we first note that almost all “nice” properties of a manifold are equivalent.

THEOREM. The following properties are equivalent for any manifold M :

- (a) Each component of M is σ -compact.
- (b) Each component of M is second countable (has a countable base for the topology).
- (c) M is metrizable.
- (d) M is paracompact.

(In particular, a compact manifold is metrizable.)

FIRST PROOF. (a) \Rightarrow (b) follows immediately from the simple proposition that a σ -compact locally second countable space is second countable.

(b) \Rightarrow (c) follows from the Urysohn metrization theorem.

(c) \Rightarrow (d) because any metric space is paracompact (Kelley, *General Topology*, pg. 160). The second proof does not rely on this difficult theorem.

(d) \Rightarrow (a) is a consequence of the following.

LEMMA. A connected, locally compact, paracompact space is σ -compact.

Proof. There is a locally finite cover of the space by open sets with compact closure. If U_0 is one of these, then \bar{U}_0 can intersect only a finite number U_1, \dots, U_{n_1} of the others. Similarly $\bar{U}_0 \cup \bar{U}_1 \cup \dots \cup \bar{U}_{n_1}$ intersects only $U_{n_1+1}, \dots, U_{n_2}$; and so on. The union

$$\bar{U}_0 \cup \dots \cup \bar{U}_{n_1} \cup \dots \cup \bar{U}_{n_2} \cup \dots = U_0 \cup \dots \cup U_{n_1} \cup \dots \cup U_{n_2} \cup \dots$$

is clearly open. It is also closed, for if x is in the closure, then x must be in the closure of a finite union of these U_i , because x has a neighborhood which intersects only finitely many. Thus x is in the union.

Since the space is connected, it equals this countable union of compact sets.

This proves the Lemma and the Theorem.

SECOND PROOF. (a) \Rightarrow (b) \Rightarrow (c) and (d) \Rightarrow (a) as before.

(c) \Rightarrow (a) is Theorem 1-2.

(a) \Rightarrow (d). Let $M = C_1 \cup C_2 \cup \dots$, where each C_i is compact. Clearly C_1 has an open neighborhood U_1 with compact closure. Then $\bar{U}_1 \cup C_2$ has an open neighborhood U_2 with compact closure. Continuing in this way, we obtain open sets U_i with \bar{U}_i compact and $\bar{U}_i \subset U_{i+1}$, whose union contains all C_i , and hence is M . It is easy to show from this that M is paracompact. ♦

It turns out that there are even 1-manifolds which are not paracompact. The construction of these examples requires the ordinal numbers, which are briefly explained here. (Ordinal numbers will not be needed for a 2-dimensional example to come later.

ORDINAL NUMBERS

Recall that an **ordering** $<$ on a set A is a relation such that

- (1) $a < b$ and $b < c$ implies $a < c$ for all $a, b, c \in A$ (transitivity)

(2) For all $a, b \in A$, one and only one of the following holds:

- (i) $a = b$
 - (ii) $a < b$
 - (iii) $b < a$ (also written $a > b$)
- (trichotomy).

An ordered set is just a pair $(A, <)$ where $<$ is an ordering on A . Two ordered sets $(A, <)$ and $(B, <)$ are order isomorphic if there is a one-one onto function $f: A \rightarrow B$ such that $a < b$ implies $f(a) < f(b)$; the map f itself is called an order isomorphism, and f^{-1} is easily seen to be an order isomorphism also.

An ordering $<$ on A is a well-ordering if every non-empty subset $B \subset A$ has a *first* element, that is, an element b such that $b \leq b'$ for all $b' \in B$. Some well-ordered sets are illustrated below; in this scheme we do not list any of the $<$ relations which are consequences of the ones already listed.

$$\begin{array}{ll} \emptyset & \\ \{0\} & \\ 0 < 1 & (A = \{0, 1\}) \\ 0 < 1 < 2 & (A = \{0, 1, 2\}) \\ 0 < 1 < 2 < 3 & \text{etc.} \\ \vdots & \\ 0 < 1 < 2 < 3 < \dots & \\ 0 < 1 < 2 < \dots < \omega & (\omega \text{ is some set } \neq 0, 1, 2, 3, \dots) \\ & (\omega + 1 \text{ is, for the present,} \\ 0 < 1 < 2 < \dots < \omega < \omega + 1 & \text{just a set distinct from} \\ & \text{those already mentioned}) \\ \vdots & \\ 0 < 1 < 2 < \dots < \omega < \omega + 1 < \omega + 2 < \dots & \\ 0 < 1 < 2 < \dots < \omega < \omega + 1 < \omega + 2 < \dots < \omega \cdot 2 & \\ \vdots & \\ 0 < 1 < 2 < \dots < \omega < \omega + 1 < \omega + 2 < \dots < \omega \cdot 2 < \omega \cdot 2 + 1 < \dots & \\ 0 < 1 < 2 < \dots < \omega < \omega + 1 < \omega + 2 < \dots < \omega \cdot 2 < \omega \cdot 2 + 1 < \dots < \omega & \\ \vdots & \\ 0 < 1 < 2 < \dots < \omega < \dots < \omega \cdot 2 < \dots < \omega \cdot 3 < \dots < \dots & \\ 0 < 1 < 2 < \dots < \omega < \dots < \omega \cdot 2 < \dots < \omega \cdot 3 < \dots < \dots < \omega^2 & \end{array}$$

Any subset of a well-ordered set is, of course, also a well-ordered set with the same ordering. In particular, a subset B of a well-ordered set A is called an (initial) segment if $b \in B$ and $a < b$ imply $a \in B$. It is easy to see that if B is a segment of A , then either $B = A$ or else there is some $a \in A$ such that

$$B = \{a' \in A : a' < a\};$$

in fact, a is the first element of $A - B$. Notice that each set on our list is a segment of the succeeding ones. It is not hard to see that no two sets on our list are order isomorphic. For example,

$$0 < 1 < \dots < \omega \quad \text{and} \quad 0 < 1 < \dots < \omega < \omega + 1$$

are not order isomorphic because the second has both a last and a next to last element, while the first does not. But there is a much more general proposition which will settle all cases at once:

1. PROPOSITION. If $B \neq A$ is a segment of A , then B is not order isomorphic to A . In fact, the only order isomorphism from B to a segment of A is the identity.

PROOF. If $f: B \rightarrow B' \subset A$ is an order isomorphism and B' is a segment of A , then for the first element b of B (and hence of A) we clearly must have $f(b) = b$. Then $f(b')$ must be b' , where b' is the second element. And so on, even for the " ω^{th} " element (the first one after the first, second, third, etc.)! The way we prove this rigorously is amazingly simple: If $f(b) \neq b$ for some $b \in B$, just consider the first element of $\{b \in B : f(b) \neq b\}$; an outright contradiction appears almost immediately. ♦

Proposition 1 has a companion, which makes the study of well-ordered sets simply delightful.

2. PROPOSITION. If $(A, <)$ and $(B, <)$ are well-ordered sets, then one is order isomorphic to a segment of the other.

PROOF. We match the first element of A with the first of B , the second with the second, ..., the " ω^{th} " with the " ω^{th} ", etc., until we run out of one set. To do this rigorously, consider order isomorphisms from segments of A onto segments of B . It is easy to show that any two such order isomorphisms agree on the smaller of their two domains (just consider the smallest element where

they don't). So all such order isomorphisms can be put together to give another, which is clearly the largest of all. If it is defined on all of A we are done. If it is not, then its range must be all of B (or we could easily extend it) and we are still done. ♦

Suppose we define a relation $<$ between well-ordered sets by stipulating that $(A, <) < (B, <)$ when $(A, <)$ is order isomorphic to a proper segment of $(B, <)$. Transitivity of $<$ is obvious, and Propositions 1 and 2 show that we almost have trichotomy. "Almost", because the condition " $(A, <) = (B, <)$ " must be replaced by " $(A, <) \text{ order isomorphic to } (B, <)$ ". To obviate this difficulty we need only work with order isomorphism classes of well-ordered sets, instead of with the well-ordered sets themselves. These order isomorphism classes are called **ordinal numbers**. They are beautiful.*

3. PROPOSITION. $<$ is a well-ordering of the ordinal numbers.

PROOF. Given a non-empty set \mathcal{A} of ordinal numbers, let $(A, <)$ be a well-ordered set representing one of its elements α . To produce a smallest element of \mathcal{A} we can obviously ignore elements $\geq \alpha$. Every element $< \alpha$ is represented by an ordered set which is order isomorphic to some proper segment of A : each of these is the segment consisting of elements of A less than some $a \in A$. Consider the least of these a 's. It determines a segment which represents some $\beta \in \mathcal{A}$. This β is the smallest element of \mathcal{A} . ♦

Notice that if α is an ordinal number, represented by a well-ordered set $(A, <)$, then the well-ordered set of all ordinals $\beta < \alpha$ has a particularly simple representation: it is order isomorphic to the set $(A, <)$! Roughly speaking: An ordinal number is order isomorphic to the set of all ordinals less than it.

If α is an ordinal number, we will denote by $\alpha + 1$ the smallest ordinal after α (if α is represented by the well-ordered set $(A, <)$, then $\alpha + 1$ is represented by a well-ordered set with just one more element, larger than all members of A). Notice that some ordinals are not of the form $\alpha + 1$ for any α ; these are called **limit ordinals**, while those of the form $\alpha + 1$ are called **successor**

*Only one feature mars the beauty of the ordinal numbers as presented here. Each ordinal number is a horribly large set; it would be much nicer to choose one specific well-ordered set from each order isomorphism class, and define these specific sets to be the ordinal numbers. There is a particularly elegant way to do this, due to von Neumann, which can be found in the Appendix to Kelley, *General Topology*.

ordinals. We will also denote some ordinals by the symbols appearing before: $0, 1, 2, 3, \dots, \omega, \omega + 1, \dots$, etc.

Our list of well-ordered sets only begins to suggest the complexity which well-ordered sets can achieve. With a little thought, one can see how the symbols $\omega^3, \omega^4, \dots$ would appear (symbols like $\omega^3 + \omega^2 \cdot 3 + \omega \cdot 4 + 6$ would be used somewhere between ω^3 and ω^4): after all these one would need

$$\omega^\omega, \omega^{\omega^\omega}, \dots$$

and after all these the symbol ε_0 pops up. After

$$\varepsilon_0^2, \varepsilon_0^3, \dots, \varepsilon_0^\omega, \dots, \varepsilon_0^{\omega^\omega}, \dots, \varepsilon_0^{\omega^{\omega^\omega}}, \dots,$$

one comes to

$$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_\omega, \dots, \varepsilon_{\omega^\omega}, \dots, \varepsilon_{\varepsilon_0}, \dots, \varepsilon_{\varepsilon_{\varepsilon_0}}, \dots;$$

and this is only the beginning!

All the well-ordered sets mentioned so far are *countable*. There are indeed an enormous number of countable well-ordered sets:

4. PROPOSITION. Let Ω be the collection of all countable ordinals (ordinals represented by a countable well-ordered set). Then Ω is uncountable.

PROOF. By Proposition 3, $(\Omega, <)$ is a well-ordered set. If it were countable, it would represent a countable ordinal $\alpha \in \Omega$. By the remark after Proposition 3, this would mean that Ω is order isomorphic to the collection of ordinals $< \alpha$, i.e., to a proper segment of itself, contradicting* Proposition 1. ♦

We have thus established the existence of an uncountable ordinal. Our specific example, represented by Ω , is clearly the first uncountable ordinal; any member of Ω is countable, and consequently has only countably many predecessors. (It is hopeless to try to “reach” Ω by continuing the listing of well-ordered sets begun above, for one would have to go uncountably far, and encounter sets with an uncountable number of degrees of complexity. A leap of faith is required.)

Although the countable ordinals exhibit uncountably many degrees of complexity, they are each simple in one way:

* By deleting the words countable and uncountable in this proof one obtains the “Burali-Forti Paradox”: the set *Ord* of all ordinal numbers is well-ordered, so it represents an ordinal $\alpha \in \text{Ord}$, and hence is order isomorphic to an initial segment of *Ord*. For a resolution of this paradox, see Kelley’s Appendix.

5. PROPOSITION. If $\alpha \in \Omega$ is a limit ordinal, then there is a sequence $\beta_1 < \beta_2 < \beta_3 < \dots < \alpha$, such that every $\beta < \alpha$ satisfies $\beta < \beta_n$ for some n (we say that $\{\beta_n\}$ is “cofinal” in α).

PROOF. Since α is countable, *all* its members can be listed (in not-necessarily increasing order) $\gamma_1, \gamma_2, \gamma_3, \dots$. Let $\beta_1 = \gamma_1$ and let β_{n+1} be the first γ in the list which comes after β_n . ♦

6. COROLLARY. If $\alpha \in \Omega$, then α is represented by some well-ordered subset of \mathbb{R} . However, no subset of \mathbb{R} is order isomorphic to Ω .

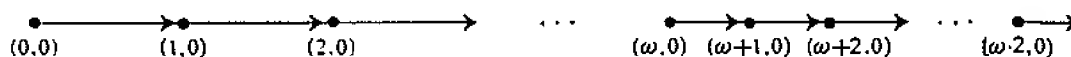
PROOF. Suppose there were one, and hence a smallest, $\alpha \in \Omega$ not represented by some subset of \mathbb{R} . It cannot happen that $\alpha = \beta + 1$, for then β would be represented by a subset of \mathbb{R} , thus also by a subset of $(-\infty, 0)$ and α could be represented by a subset of \mathbb{R} . So by Proposition 5, there is a sequence $\beta_1 < \beta_2 < \beta_3 < \dots < \alpha$ cofinal in α . Then β_i is represented by a subset of $(-\infty, i)$, and we can easily arrange that the subset representing β_i is a segment of the subset representing β_j for $i < j$. The union of all these sets would then represent α , a contradiction.

If a subset of \mathbb{R} were order isomorphic to Ω , then there would be uncountably many disjoint intervals in \mathbb{R} , namely those between the points representing α and $\alpha + 1$ for all $\alpha \in \Omega$. This is impossible. ♦

The first example of a non-metrizable manifold is defined in terms of Ω . Consider $\Omega \times [0, 1)$, with the order $<$ defined as follows:

$$(\alpha, s) < (\beta, t) \quad \text{if } \alpha < \beta \text{ or if } \alpha = \beta \text{ and } s < t.$$

This can be pictured as follows:



The set $\Omega \times [0, 1)$ with the order topology (a subbase consists of sets of the form $\{x : x < x_0\}$ and $\{x : x > x_0\}$) is called the closed long ray (with “origin” $(0, 0)$), and $L^+ = \Omega \times [0, 1) - \{(0, 0)\}$ is the (open) long ray. The disjoint union of two copies of the closed long ray with their origins identified is the long line L . To distinguish L^+ and L , the names “half-long line” and “long line” may also be used. The Corollary to Proposition 5 implies easily that the long ray and the long line are 1-dimensional manifolds; aside from the line and the circle, there are no other connected 1-manifolds.

Quite a few new 2-manifolds can now be constructed:

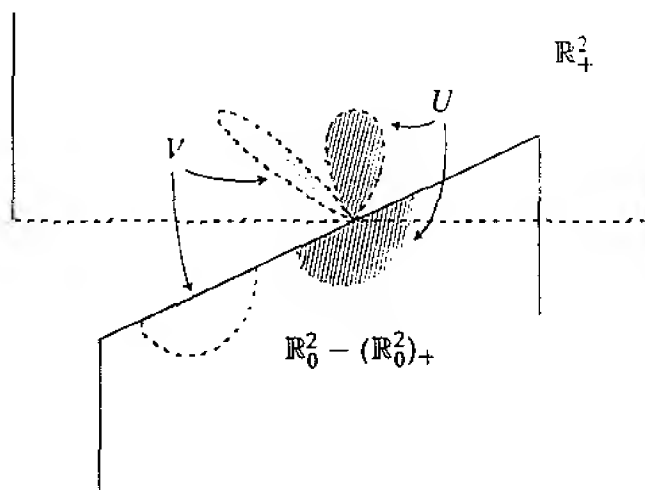
$$\begin{array}{ll}
 L^+ \times S^1 & \text{(half-long cylinder),} \\
 L^+ \times \mathbb{R} & \text{(half-long strip),} \\
 L \times L & \text{(big plane),} \\
 L^+ \times L^+ & \text{(big quadrant).}
 \end{array}
 \qquad
 \begin{array}{ll}
 L \times S^1 & \text{(long cylinder),} \\
 L \times \mathbb{R} & \text{(long strip),} \\
 L \times L^+ & \text{(big half-plane),}
 \end{array}$$

Identifying all points $((0, 0), \theta)$ in the product of the closed long ray and S^1 produces another 2-manifold, which might be called the "big disc".

There is another way of producing a non-metrizable 2-manifold which does not use Ω at all. We begin with the open upper half-plane $\mathbb{R}_+^2 = \{(x, y) \in \mathbb{R}^2 : y > 0\}$ and another copy $\mathbb{R}^2 \times \{0\}$ of the plane; we will denote this set by \mathbb{R}_0^2 , and denote the point $(x, y, 0)$ by $(x, y)_0$. Define a map $f_0: (\mathbb{R}_0^2)_+ \rightarrow \mathbb{R}_+^2$ by

$$f_0((x, y)_0) = (xy, y).$$


Consider the disjoint union of \mathbb{R}_+^2 and \mathbb{R}_0^2 , with $p \in (\mathbb{R}_0^2)_+$ and $f_0(p) \in \mathbb{R}_+^2$ identified. This is a Hausdorff manifold; the following diagram shows two open sets homeomorphic to \mathbb{R}^2 . The manifold itself is, in fact, homeomorphic



to \mathbb{R}^2 ; we could have thrown away \mathbb{R}_+^2 to begin with since it is identified by a homeomorphism with $(\mathbb{R}_0^2)_+$.

But consider now, for each $a \in \mathbb{R}$, another copy of \mathbb{R}^2 , say $\mathbb{R}^2 \times \{a\}$, which we will denote by \mathbb{R}_a^2 . Define $f_a: (\mathbb{R}_a^2)_+ \rightarrow \mathbb{R}_+^2$ by

$$f_a((x, y)_a) = (a + xy, y).$$


In the disjoint union of \mathbb{R}_+^2 and *all* \mathbb{R}_a^2 , $a \in \mathbb{R}$ we wish to identify each $p \in (\mathbb{R}_a^2)_+$ with $f_a(p) \in \mathbb{R}_+^2$. We may dispense with \mathbb{R}_+^2 completely, and in the disjoint union of all \mathbb{R}_a^2 identify each $(x, y)_a$ and $(x', y')_b$ for which $y = y' > 0$ and $xy + a = x'y' + b$. The equivalence classes, of course, are a space homeomorphic to \mathbb{R}_+^2 , so we will consider \mathbb{R}_+^2 a subset of the resulting space. This space is still a Hausdorff manifold, but it cannot be second countable, for it has an uncountable discrete subset, namely the set $\{(0, 0)_a\}$. This manifold, the **Prüfer manifold**, and related manifolds, have some very strange properties, developed in the problems.

PROBLEMS

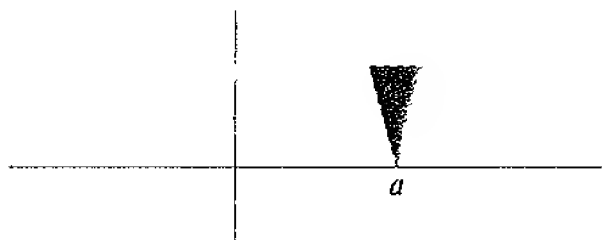
1. (a) A well-ordered set cannot contain a decreasing infinite sequence $x_1 > x_2 > x_3 > \cdots$.
- (b) If we denote $(\alpha + 1) + 1$ by $\alpha + 2$, $(\alpha + 2) + 1$ by $\alpha + 3$, etc., then any α equals $\beta + n$ for a unique limit ordinal β and integer $n \geq 0$. (Thus one can define *even* and *odd* ordinals.)
2. Let c be a “choice function”, i.e., $c(A)$ is defined for each set $A \neq \emptyset$, and $c(A) \in A$ for all A . Given a set X , a well-ordering $<$ on a subset Y of X will be called “distinguished” if for all $y \in Y$,

$$y = c(Y - \{y' \in Y : y' < y\}).$$

- (a) Show that of any two distinguished well-orderings, one is an extension of the other.
- (b) Show that there is a well-ordering on X . (Zorn's Lemma may be deduced from this fact fairly easily.)
- (c) Given two sets, show that one of them is equivalent to (can be put in one-one correspondence with) a subset of the other.
- (d) Show that on any infinite set there is a well-ordering which represents a limit ordinal.
- (e) From (d), and Problem 1, show that if X and Y are disjoint equivalent infinite sets, then $X \cup Y$ is equivalent to Y .

3. (a) L^+ and L are not metrizable.
- (b) If $x_1 \leq x_2 \leq x_3 \leq \cdots$ is a sequence in L^+ , then $\{x_n\}$ converges to some point. Consequently, any sequence has a convergent subsequence (but L^+ is not compact!).

- (c) If $\{x_n\}$ and $\{y_n\}$ are sequences in L^+ with $x_n \leq y_n \leq x_{n+1}$ for all n , then both sequences converge to the same point.
- (d) L^+ (and also L) are normal. (Use (c)).
- (e) More generally, any order topology is normal (completely different proof).
- (f) If $f: L^+ \rightarrow \mathbb{R}$ is continuous, and $r > s$, then one of the sets $f^{-1}((-\infty, s])$ and $f^{-1}([r, \infty))$ is countable.
- (g) If $f: L^+ \rightarrow \mathbb{R}$ is continuous, then f is eventually constant.
4. (a) L^+ is not contractible. *Hint:* Given $H: L^+ \times [0, 1] \rightarrow L^+$ with $H(x, 0) = x$ for all x , show that for every t we have $\{H(x, t)\} = L^+$.
- (b) $\pi_1(L^+) = \pi_1(L) = 0$. Similarly for $L^+ \times \mathbb{R}$, $L \times \mathbb{R}$, $L \times L$, $L \times L^+$, $L^+ \times L^+$.
- (c) $\pi_1(L^+ \times S^1) = \pi_1(L \times S^1) = \mathbb{Z}$.
5. (a) L^+ and L are not homeomorphic. *Hint:* Imitating Problem 1-19, define “paracompact ends”.
- (b) $L^+ \times \mathbb{R}$ and $L \times \mathbb{R}$ are not homeomorphic; $L^+ \times S^1$ and $L \times S^1$ are not homeomorphic.
- (c) Of the 2-manifolds constructed from L^+ or L with $\pi_1 = 0$ and one paracompact end, only $L^+ \times \mathbb{R}$ has the homotopy type of L^+ .
- (d) The Stone-Čech compactifications of $L \times L$, $L^+ \times L$, $L^+ \times L^+$, and the big disc are all distinct. (Using Problem 3(g), one can explicitly construct these Stone-Čech compactifications).
6. (a) Show that the Prüfer manifold P is Hausdorff.
- (b) P does not have a countable dense subset.
- (c) Let U be an open set in \mathbb{R}_+^2 which is the union of “wedges” centered at $(a, 0)$ for every irrational a . Show that U includes a whole rectangle of the form



$(a, b) \times (0, \varepsilon)$. *Hint:* Let $A_n = \{a : \text{the wedge centered at } a \text{ has width } \geq 1/n\}$. Since $\mathbb{R} = \mathbb{Q} \cup \bigcup_n A_n$, some A_n is not nowhere dense.

(d) Let $C_1, C_2 \subset P$ be

$$C_1 = \{(0, 0)_a : a \text{ irrational}\}$$

$$C_2 = \{(0, 0)_a : a \text{ rational}\}.$$

Show that C_1 and C_2 are closed, but that they are not contained in disjoint open sets.

(e) Define $H: P \times [0, 1] \rightarrow P$ by

$$H((x, y)_a, s) = \begin{cases} \left(x \sqrt{\frac{1-s+sy}{1+sy}}, y \sqrt{\frac{1+sy}{1-s+sy}} \right)_a & \text{if } y > 0 \\ \left(x \sqrt{1-s^2}, y \sqrt{1-s^2} \right)_a & \text{if } y \leq 0. \end{cases}$$

Show that H is well-defined and that $H(p, 1) \in \mathbb{R}_+^2 \cup \{(0, 0)_a\}$ for all $p \in P$. Conclude that P is contractible.

(f) $P - \{(x, y)_a : y < 0\}$ is a manifold-with-boundary P' , whose boundary is a disjoint union of uncountably many copies of \mathbb{R} .

(g) The disjoint union of two copies of P' , with corresponding points on the boundary identified, is a manifold which is not metrizable, but which has a countable dense subset. Its fundamental group is uncountable.

7. It is known that every second countable contractible 2-manifold is S^2 or \mathbb{R}^2 . Hence the result of constructing the Prüfer manifold using only copies \mathbb{R}_a^2 for rational a must be homeomorphic to \mathbb{R}^2 . Describe a homeomorphism of this manifold onto \mathbb{R}^2 .

8. Let M be a connected Hausdorff manifold which is not a point.

(a) If $A \subset M$ has cardinality \mathfrak{c} (the cardinality of \mathbb{R}), then the closure \bar{A} has cardinality \mathfrak{c} .

(b) If $C \subset M$ is closed and has cardinality \mathfrak{c} , then C has an open neighborhood with cardinality \mathfrak{c} .

(c) Let $p \in M$. There is a function $f: \Omega \rightarrow (\text{set of subsets of } M)$ such that $f(\alpha)$ has cardinality \mathfrak{c} for all $\alpha \in \Omega$, and such that

$$f(0) = \{p\}$$

$$f(\alpha) \text{ is an open neighborhood of the closure of } \bigcup_{\beta < \alpha} f(\beta).$$

(Consider functions defined on initial segments of Ω with these same properties, and apply Zorn's Lemma. Alternatively, one can require $f(\alpha)$ to be the result of applying the choice function to the set of all open neighborhoods of the closure

of $\bigcup_{\beta < \alpha} f(\beta)$ with cardinality \mathfrak{c} . Then there is a unique f with the required properties. This is an example of defining a function by “transfinite induction”.)

(d) A function $f: \Omega \rightarrow (\text{set of subsets of } [0, 1])$ with the properties of the function in part (c) is eventually constant.

(e) M has cardinality \mathfrak{c} . (Given $p' \in M$, consider an arc from p to p' .)

9. (a) A connected 1-manifold whose topology is the order topology for some order, is homeomorphic to either the real line, the long line, or the half-long line.

(b) Every 1-manifold M contains a maximal open submanifold N whose topology is the order topology for some order.

(c) If M is connected and $N \neq M$, then M is homeomorphic to S^1 .

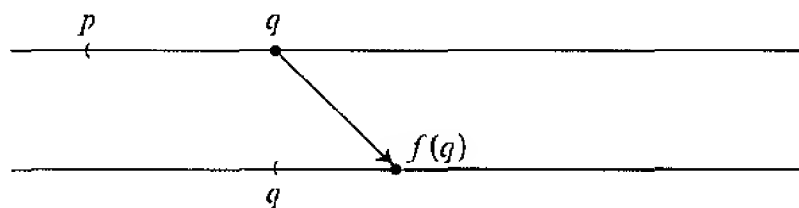
CHAPTER 2

The long ray L^+ can be given a C^∞ structure, and even a C^ω structure. To see this we need the result of Problem 9-24—any C^∞ [or C^ω] structure on a manifold M homeomorphic to \mathbb{R} is diffeomorphic to \mathbb{R} with the usual structure. This implies that it is also diffeomorphic to $(0, 1)$, and consequently that the structure on M can be extended if M is a proper subset of L^+ . An easy application of Zorn's Lemma then shows that C^∞ and C^ω structures exist on L^+ .

I do not know whether all C^∞ structures on L^+ are diffeomorphic. It is known that there are uncountably many inequivalent C^ω structures on L^+ . If $p \in L^+$, and L^+_p denotes all points $\leq p$, then $L^+ - L^+_p$ is clearly homeomorphic to L^+ . If \mathcal{O} is a C^ω structure for L^+ , then it yields a C^ω structure for $L^+ - L^+_p$, and hence for L^+ . These are all distinct, in other words, there is no C^ω map

$$f: L^+ - L^+_p \rightarrow L^+ - L^+_q \quad q > p$$

with a C^ω inverse. In fact, we must have $f(q) > q$, and then it is easy to see



that we must also have $f(f(q)) > f(q)$, $f(f(f(q))) > f(f(q))$, etc. The increasing sequence $q, f(q), f(f(q)), \dots$ has a limit point $x_0 \in L^+ - L^+_p$, and $f(x_0) = x_0$. Now f cannot be the identity on all points $> x_0$ (for then it would be the identity everywhere, since it is C^ω). So for some $q_1 > x_0$ we have $f(q_1) \neq q_1$; we can assume $f(q_1) > q_1$, since we can consider f^{-1} in the contrary case. Reasoning as before, we obtain $x_1 > x_0$ with $f(x_1) = x_1$. Continuing in this way, we obtain $x_0 < x_1 < x_2 < \dots$ with $f(x_n) = x_n$. This sequence has a limit in $L^+ - L^+_p$, but this implies that $f(x) = x$ for all x , a contradiction.

A C^ω structure exists on the Prüfer manifold; this follows immediately from the fact that the maps f_a , used for identifying points in various $(\mathbb{R}^2_0)_+$ with points in \mathbb{R}^2_+ , are all C^ω . I do not know whether every 2-manifold has a C^∞ structure.

Using the C^ω structure on L^+ , we can get a C^ω structure on $L^+ \times L^+$. However, the method used for obtaining a C^ω structure on L^+ will not yield a *complex analytic structure* on $L^+ \times L^+$; the problem is that a complex analytic structure on \mathbb{R}^2 may be conformally equivalent to the disc, and hence extendable, but it

may also be equivalent to the complex plane, and not extendable. In fact, it is a classical theorem of Rado that every Riemannian surface (2-manifold with a complex analytic structure) is second countable. On the other hand, a modification of the Prüfer manifold yields a non-metrizable manifold of complex dimension 2. References to these matters are to be found in

Calabi and Rosenlicht, *Complex Analytic Manifolds without Countable Base*, Proc. Amer. Math. Soc. 4 (1953), pp. 335–340.

H. Kneser, *Analytische Struktur und Abzählbarkeit*, Ann. Acad. Sic. Fennicae Series A, I 251/5 (1958), pp. 1–8.

PROBLEMS

10. Prove that for $q > p$ there is no non-constant C^ω map $f: L^+ - L^+_p \rightarrow L^+ - L^+_q$.
11. Let (Y, ρ) be a metric space and let $f: X \rightarrow Y$ be a continuous locally one-one map, where X is Hausdorff, connected, locally connected, and locally compact.
 - (a) Every two points $x, y \in X$ are contained in a compact connected $C \subset X$.
 - (b) Let $d(x, y)$ be the greatest lower bound of the diameters of $f(C)$ (in the ρ -metric) for all compact connected C containing x and y . Show that d is a metric on X which gives the same topology for X .
12. Of the various manifolds mentioned in the previous section, try to determine which can be immersed in which.

CHAPTER 6

Problem A-6(g) describes a non-paracompact 2-manifold in which two open half-planes are a dense set. We will now describe a 3-dimensional version with a twist.

Let $A = \{(x, y, z) \in \mathbb{R}^3 : y \neq 0\}$, and for each $a \in \mathbb{R}$ let \mathbb{R}_a^3 be a copy of \mathbb{R}^3 , points in \mathbb{R}_a^3 being denoted by $(x, y, z)_a$. In the disjoint union of A and all \mathbb{R}_a^3 , $a \in \mathbb{R}$ we identify

$$\begin{aligned} (x, y, z)_a & \text{ for } y > 0 & \text{ with } & (a + yx, y, z + a) \\ (x, y, z)_a & \text{ for } y < 0 & \text{ with } & (a + yx, y, z - a). \end{aligned}$$

The equivalence classes form a 3-dimensional Hausdorff manifold M . On this manifold there is an obvious function “ z ”, and the sets $z = \text{constant}$ form a foliation of M by a 2-dimensional manifold N . The remarkable fact about this 2-dimensional manifold N is that it is connected. For, the set of points $(x, y, c) \in A$ with $y > 0$ is identified with the set of points $(x, y, c - a)_a \in \mathbb{R}_a^3$ with $y > 0$. Now the folium containing $\{(x, y, c - a)_a\}$ contains the points $(x, y, c - a)_a$ with $y < 0$, and these are identified with the set of points $(x, y, c - 2a) \in A$ with $y < 0$. Since we can choose $a = c/2$, we see that all leaves of the foliation are the same as the leaf containing $\{(x, y, 0) : y < 0\} \subset A$.

This example is due to M. Kneser, *Beispiel einer dimensionserhöhenden analytischen Abbildung zwischen überabzählbaren Mannigfaltigkeiten*. Archiv. Math. 11 (1960), pp. 280–281.

CHAPTERS 7, 9, 10

1. We have seen that any paracompact C^∞ manifold has a Riemannian metric. The converse also holds, since a Riemannian metric determines an ordinary metric.

2. Problem A-11 implies that a manifold N immersed in a paracompact manifold M is paracompact, but a much easier proof is now available: Let (\cdot, \cdot) be a Riemannian metric on M ; if $f: N \rightarrow M$ is an immersion, then N has the Riemannian metric $f^*(\cdot, \cdot)$.

We can now dispense with the argument in the proof of Theorem 6-6 which was used to show that each folium of a distribution on a metrizable manifold is also metrizable, for the folium is a submanifold, and hence paracompact.

3. Since there is no Riemannian metric on a non-paracompact manifold M , the tangent bundle TM cannot be trivial. Thus the tangent bundle of the long line is not trivial, nor is the tangent bundle of the Prüfer manifold, even though the Prüfer manifold is contractible. (On the other hand, a basic result about bundles says that a bundle over a paracompact contractible space is trivial. Compare pg. V.272.)

4. The tangent bundle of the long line L is clearly orientable, so there *cannot* be a nowhere zero 1-form ω on L , for ω and the orientation would determine a nowhere zero vector field, contradicting the fact that the tangent bundle is not trivial. Thus, Theorem 7-9 fails for L . Notice also that if M is non-paracompact, then TM is definitely not equivalent to T^*M , since an equivalence would determine a Riemannian metric. So there are at least two inequivalent non-trivial bundles over M .

5. Although the results in the Addendum to Chapter 9 can be extended to closed, not necessarily compact, submanifolds, they cannot be extended to non-paracompact manifolds, as can be seen by considering the 0-dimensional submanifold $\{(0,0)_a\}$ of the Prüfer manifold.

6. A Lie group is automatically paracompact, since its tangent bundle is trivial. More generally, a locally compact connected topological group is σ -compact (Problem 10-4).

7. It is not clear that a non-paracompact manifold cannot have an indefinite metric (a non-degenerate inner product on each tangent space). This will be proved in Volume II (Chapter 8, Addendum 1).

PROBLEM

13. Is there a nowhere zero 2-form on the various non-paracompact 2-manifolds which have been described?

NOTATION INDEX

CHAPTER 1

$\mathcal{E}(X)$	23
\mathbb{H}^n	19
M^n	4
\mathbb{P}^2	11
\mathbb{P}^n	19
\mathbb{R}^n	1
S^1	6
S^n	7
∂M	19

CHAPTER 2

A^t	61
C^r	34
C^0	34
C^∞	28
C^ω	34
$D_i f(a)$	35
$\mathrm{GL}(n, \mathbb{R})$	61
$O(n)$	61
$R(n)$	62
$(\mathbb{R}^n, \mathcal{U})$	29
$\mathrm{SL}(n, \mathbb{R})$	61
$\mathrm{SO}(n)$	62
(x, U)	28
$\frac{\partial f}{\partial x^i}(p), \frac{\partial f}{\partial x^i} \Big _p$	35
$\frac{\partial}{\partial x^i} \Big _p$	39
$\frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}$	36

CHAPTER 3

Df	65
$\varepsilon^n(X)$	72
\mathcal{F}	83
f_\star	65, 75
$f_{\star p}$	65
$f^\star(\xi)$	101

M_p	76
$(M, i)_p$	68
\mathbb{R}^n_p	64
TM	75
$T(M, i)$	68
$T\mathbb{R}^n$	64
T^\star	103
X_p	82
\bar{X}	83
\dot{x}^i	81
$[x, v]_p$	76
v_p	64
$v(f)$	80
$[v_1, \dots, v_n]$	84
ξA	72
$\xi \oplus \eta$	101
$\xi_1 \times \xi_2$	102
$\frac{dc}{dt}$	81
$\frac{dc}{dt} \Big _{t_0}$	80
$\frac{\partial}{\partial x^i}$	83

CHAPTER 4

df	109
dx^i	110
$\mathrm{End}(V)$	121
f^\star	107, 116, 119
f_p^\star	113
$\mathrm{Hom}(V, W)$	131
$T^\star M$	109
$T \otimes S$	116
$\mathcal{T}^k(V)$	116
$\mathcal{T}^k(\xi)$	117
$\mathcal{T}_k(V)$	120
$\mathcal{T}_1^1(V)$	121
$\mathcal{T}_1^1(\xi)$	121
$\mathcal{T}_l^k(V)$	122

$\mathcal{T}_l^k(\xi)$	123	$\mathcal{T}_l^{k[m]}(V)$	231
V^*	107	$\mathcal{T}_{l[m]}^k(V)$	231
V	117	$\mathcal{T}_l^{k[n;w]}(V)$	231
v^*	107	$\mathcal{T}_{l[n;w]}^k(V)$	231
v^{**}	107	$\mathcal{T}^0(V)$	201
$\delta_{i_1 \dots i_l}^{j_1 \dots j_l}$	129	$v \perp \omega$	227
$\varepsilon_{i_1 \dots i_n}$	134	ϕ_l	206
$\varepsilon^{i_1 \dots i_n}$	134	$\sigma \cdot (v_1, \dots, v_k)$	202
ξ^*	108	$\sigma \bullet (v_1, \dots, v_k)$	227
$\omega(X)$	109	$\Omega(M)$	215
\mathfrak{h}_V	107	$\Omega^k(V)$	201
CHAPTER 5		$\Omega^0(V)$	201
$ A $	171	\wedge	203
c''	161	CHAPTER 8	
$\exp A$	171	$B^k(M)$	263
$L_X A$	174	$B_c^k(M)$	268
$L_X f$	150	$c_{(i, \alpha)}$	250
$L_X Y$	150	$c_{R, n}$	284
$L_X \omega$	150	$\deg f$	275
$o(t^2)$	177	$ dx^1 \wedge \dots \wedge dx^n $	258
$[X, Y]$	153	$d\theta$	252
$\alpha_x(t) = \alpha(t, x)$	143	$d\Theta_n$	290
$(\alpha_* X)_q$	135	$d\Theta_{(a, b, c)}$	297
ϕ_t	144	f'	292
CHAPTER 7		f^*	274
Ab	202	$H^k(M)$	263
$\overline{\text{Ab}}$	205	$H_c^k(M)$	268
$\text{curl } X$	238	j^k	246
$\text{div } X$	238	$j_{(i, \alpha)}^n$	249
$d\theta$	219, 235	$\ell(f, g)$	296
$d\omega$	210, 213, 215, 234	M_i	283
$\text{grad } f$	237	m_i	283
$/\omega$	224	$\ P\ $	239
$\mathfrak{L}(\Delta)$	215	r	264
$i_p \omega$	227	$\text{sign}_p f$	275
$L_X \omega$	234	$w(p)$	293
S_k	202	$Z^k(M)$	263

$Z_c^k(M)$	268	\mathfrak{s}	313
Δ_n	285	\sinh	356
θ	291	\tanh	356
\mathfrak{t}	264, 291	W^\perp	349
\mathfrak{o}	264	$\dot{x}^i(v)$	335
σ^i	264	\tilde{x}^i	335
$[\omega]$	263	$\tilde{\alpha}(u)$	318
$\partial\epsilon$	248, 285	δJ	319
$\partial_t\epsilon$	285	$\delta\lambda$	314
$\partial\mu$	260	Γ	353
$\int_c f$	246	Γ_{ij}^k	328
$\int_c f dx + g dy$	239	(\cdot, \cdot)	301
$\int_c \omega$	243, 245, 246, 248.	$(\cdot, \cdot)_c$	315
$\int_M \omega$	257, 259, 288	$(\cdot, \cdot)_F$	308
$\#f^{-1}(q)$	294	$(\cdot, \cdot)_i$	301
$[0, 1]^0$	246	$(\cdot, \cdot)^k$	349
\cup	299	$(\cdot, \cdot)^*$	305
\times	299	$\langle \cdot, \cdot \rangle$	301
\wedge	266	$\langle \langle \cdot, \cdot \rangle \rangle$	367
		$ $	303
		$ $	303
		$ _c$	315

CHAPTER 9

\cosh	322, 356
\cosh^{-1}	356
$d(p, q)$	314
ds	314
dV	311
$Euc(V)$	309
$Euc(\xi)$	309
$E(\gamma)$	324
\exp	334
$f^*(\cdot, \cdot)$	302
(g^{ij})	306
$[ij, l]$	326
L_a^b	312
M_p^\perp	344

CHAPTER 10

$ A $	384
A^U	372
$\text{Ad}(a)$	409
$\text{ad } X$	410
$\text{Aut}(\mathfrak{g})$	409
C_{ij}^k	396
$d\omega$	402, 404
$E(n)$	373
$\text{End}(\mathfrak{g})$	410
\exp	385
$\exp(A)$	385
$\text{GL}(n, \mathbb{R})$	372
$G^{\mathfrak{a}}$	407
$\mathfrak{g}^{\mathfrak{a}}$	407
$\mathfrak{gl}(n, \mathbb{R})$	376

I_a	401
L_a	374
L_a	379
$\mathcal{L}(G)$	376
$O(t^3)$	388
$\mathfrak{o}(n)$	376
P	411
P^{-1}	411
R_a	374
$SO(n)$	373
$\tilde{\lambda}$	376
$\tilde{\lambda}$	379
ϕ_*	380
ψ	395
$\rho(\omega \wedge \eta)$	403, 410
ω (natural \mathfrak{g} -valued 1-form,	403
$[\cdot, \cdot]$	376
$[\eta \wedge \lambda]$	403
$\int_G f \sigma^n$	400
$\int_G f$	400
$\int_G f(a) da$	400
$\sum_{k=1}^d \omega^i \cdot u_i$	403

CHAPTER 11

$C^k(M)$	419
f_X	446
$\mathcal{G}^k(N)$	432
$M \# N$	453
PD	439
Rh	457
\mathcal{U}	442
Δ_*	426
δ	423
ρ	435
$\chi(M)$	428
$\chi(\xi)$	445
\cup	439

APPENDIX A

Ord	464
$\alpha + 1$	463
ε_0	464
Ω	464
ω	461
$<$	461
$<$	463

INDEX

- Abelian Lie algebra, 376, 382, 395
- Adams, J. F., 100
- Adjoint T^* of a linear transformation T , 103
- Ado, I. D., 380
- Alexander's Horned Sphere, 55
- Algebra. Fundamental Theorem of, 285, 293
- Algebraic inequalities, principle of irrelevance of, 233
- Alternating
 - covariant tensor field, 207
 - multilinear function, 201
- Alternation, 202
- Analytic manifold, 34
- Annihilator, 228
- Annulus, 8
- Antipodal
 - map, 278
 - point, 11
- Arlength, 312
 - function, 313, 332
- Arcwise connected, 20
 - subgroup of a Lie group, 409
- Area, generalized, 246
- Associated
 - disc bundle, 451
 - sphere bundle, 451
- Atlas, 28
 - maximal, 29
- Auslander, L., 106

- Banach space, 145
- Base space, 71
- Basis
 - dual, 107
 - for M_p^* , 208
 - for $\Omega^k(p)$, 208
- Belongs to a distribution, 191
- Besicovitch, A. S., 179
- Big
 - disc, 466
 - half-plane, 466
 - plane, 466
 - quadrant, 466
- Bi-invariant metric, 401
- Boundary, 19, 248, 252
- Bounded manifold, 19
- Boy's Surface, 60
- Bracket, 154
 - in $\mathfrak{gl}(n, \mathbb{R})$, 378
 - in $\mathfrak{o}(n, \mathbb{R})$, 379
- Bundle
 - cotangent, 109
 - dual, 108
 - fibre, 309
 - induced, 101
 - map, 73
 - n -plane, 71
 - normal, 344
 - of contravariant tensors, 120
 - of covariant tensors, 117
 - tangent, 77
 - trivial, 72, 210
 - vector, 71
- Burali-Forti Paradox, 464

- Calabi, E., 472
- Calculus of variations, 316
- Cartan, Élie, 39, 348, 360
- Cartan's Lemma, 230
- Cauchy-Riemann equations, 200
- Cayley numbers, 100
- Chain, 248, 285
- Chain Rule, 35, 38
- Change, infinitely small, 111
- Chart, 28
- Choice, 283
- Choice function, 467
- Circle, 6
- Closed
 - form, 218, 252
 - geodesic, 367
 - half-space, 19
 - long ray, 465
 - manifold, 19
 - subgroup of a Lie group, 391
 - submanifold, 49

- Closed (*continued*,
 - up to first order, 160
- Cofinal, 465
- Cohomology, 419
 - de Rham, 263
 - group of M with real coefficients, 263
 - of a complex, 421
- Commutative diagram, 65, 420
- Cominutative Lie algebra, 376
- Complete, geodesically, 341
- Complex, 421
 - analytic structure, 471
 - numbers of norm 1, 373
- Conjugate, 358
- Constants of structure, 396
- Continuous homomorphism, 387
- Contractible, 220, 225, 236
- Contraction, 121, 139, 227
 - Lemma, 139
- Contravariant
 - functor, 130
 - tensor field, 120
 - vector field, 113
- Convex
 - geodesically, 363
 - polyhedron, 429
- Coordinate lines, 159
- Coordinate system, 28, 158
- Coordinates, 28
- Cotangent bundle, 109
- Covariant
 - functor, 130
 - tensor field, 117
 - vector field, 113
- Cover
 - locally finite, 50
 - point-finite, 60
 - refinement of, 50
- Cramer's Rule, 372
- Critical point, 40
 - in the calculus of variations, 320
- Critical value, 40
- Cross section, 227
- Cross-cap, 14
- Cross-product, 299
- Cube, singular, 246
- Cup product, 299, 439
- Curl, 238
- Cylinder, 8
- C^1 manifold, 34
- C^0 manifold, 34
- C^∞
 - distribution, 179
 - form, 207
 - function, 32
 - manifold, 29
 - manifold-with-boundary, 32
 - Riemannian metric, 308
 - structure on TM , 82
- C^∞ -related, 28
- Darboux
 - integrable, 283
 - integral, 283
- Darboux's Theorem, 284
- Debauch of indices, 39, 123
- Decomposable, 228
- Definition, invariant, 214
- Deformation retraction, 279
- Degenerate, 286
- Degree, 275
 - mod 2, 295
- Density
 - even scalar, 133, 209
 - odd scalar, 133, 259
 - relative scalar, 231
 - scalar, 133
- Derivation, 39, 78
 - of a ring, 83
- Derived set, 25
- Descartes-Euler Theorem, 429
- Determinant, 232
- Diffeomorphic, 30
- Diffeomorphism, 30
 - one-parameter group of, 148
- Differentiable, 27, 28, 31, 32
 - at a point, 31
 - manifold, 29
 - structure, 30
 - on the long line, 471
 - on \mathbb{P}^n , 32

- Differentiable (*continued*)
 - (structure *continued*)
 - on \mathbb{R}^n , 29
 - on S^n , 30
- Differential, 210
 - equation, 136, 164
 - depending on parameters, 169
 - linear, 165
 - forms, 201
 - of a function, 109
- Dimension, 4
- Direct sum, 421
- Disc bundle, associated, 451
- Discriminant, 233
- Disjoint union, 4, 20
- Distribution, 179, 181
 - ideal of, 215
 - on torus, 180
- Divergence, 238
 - Theorem, 352
- Domain, 3
- Du Bois Reymond's Lemma, 355
- Dual
 - basis, 107
 - space, 107
 - vector bundle, 108

- Einstein summation convention, 39
- Elements of norm 1, 308
- Elliptical non-Euclidean geometry, 367
- Embedding, 49
- End, 23
 - paracompact, 466
- Endomorphism, 121
- Energy, 324
- Envelope, 358
- Equations depending on parameters, 169
- Equations of structure, 404
- Equivalence (of vector bundles), 72
 - weak, 96
- Euclidean
 - metric, 305, 315
 - motion, 374
 - n -space, 1

- Euler, 429
 - characteristic, 428
 - class, 445
- Euler's Equation, 320
- Even
 - ordinal, 467
 - relative scalar, 231
 - relative tensor, 134, 231
 - scalar density, 133, 209
- Exact
 - form, 218
 - sequence, 419, 422
 - of a pair, 433
 - of vector bundles, 103
- Exponential map, 334, 385
- Exponential of matrices, 384
- Extension, 432
- Extremal, 320

- Faith, leap of, 464
- Fibre, 64, 68, 71
- Finite
 - characteristic, 205
 - type, 438
- First element, 461
- First variation, 319, 327
- Five Lemma, 440
- Fixed point, 139
- Foliation, 194
- Folium, 194
- Force field, 240
- Form, 207
 - differential, 201
 - left invariant, 374
 - right invariant, 400
- f -related, 190
- Frobenius Integrability Theorem, 192, 215
- Fubini's theorem, 254
- Functor, 130
- Functorites, 89
- Fundamental Theorem of Algebra, 285, 293
- Fundamental Theorem of Calculus, 254

- Gauss's Lemma, 337
- General linear group, 61, 372
- Generalized area, 246
- Geodesic, 333
 - closed, 367
 - reversing map, 401
- Geodesically complete, 341
- Geodesically convex, 363
- Geodesy, 333
- Germ of k -forms, 432
- Global theory of integral manifolds, 194
- Gradient, 237
- Gram-Schmidt orthonormalization process, 304
- Grok, 84
- Group
 - Lie, 371
 - matrix, 372
 - opposite, 407
 - orthogonal, 372
 - topological, 371
- Guillemin, V. W., 106

- Hahn-Banach theorem, 145
- Hair, 69
- Half-long
 - cylinder, 466
 - line, 465
 - strip, 466
- Half-space, 19
- Handle, 8
- Hardy, G. H., 179
- Has one end, 23
- Heinelein, Robert A., 84
- Hausdorff, 459
- Homogeneous, 7
- Homomorphism
 - continuous, 387
 - of Lie algebras, 380
- Homotopic, 104, 277
- Homotopy, 104, 277
- Hopf, H., 342, 450
- Hopf-Rinow-de Rham Theorem, 342

- Hyperbolic
 - cosine, 356
 - sine, 356
 - tangent, 356

- Ideal of a Lie algebra, 410
- Identification, 10
- Imbedding, 49
 - topological, 14
- Immersed submanifold, 47
- Immersion, 46
 - topological, 14, 46
- Implicit function theorem, 60
- Indefinite metric, 350
- Independent infinitesimals, 314
- Index of inner product, 349
- Index of vector field
 - on a manifold, 447
 - on \mathbb{R}^n , 446
- Indices
 - debauch of, 39, 123
 - raising and lowering, 351
- Induced
 - bundle, 101
 - orientation, 260
- Inequalities, principle of irrelevance of algebraic, 233
- Inertia, Sylvester's Law of, 349
- Infinite volume, 312
- Infinitely small change, 111
- Infinitely small displacements, 314
- Infinitesimal generator, 148
- Infinitesimals, independent, 314
- Initial conditions, 136
 - of integral curve, 136
- Initial segment, 462
- Inner product, 227, 301
 - preserving, 304, 372
 - usual, 301
- Inside, 21
- Integrability conditions, 189
- Integrable distribution, 192
- Integrable function
 - Darboux, 283
 - Riemann, 283

- Integral
 - curve, 136
 - Darboux, 283
 - line, 239, 243
 - manifold, 179, 181
 - maximal, 194
 - of a differential equation, 136
 - Riemann, 283
 - surface, 245
- Integration, 136, 226, 239
- Invariance of Domain, 3
- Invariant, 128, 232
 - definition, 214
- Irrelevance of algebraic inequalities, principle of, 233
- Isometry, 340
- Isomorphic Lie groups, locally, 382
- Isomorphism, natural, 108
- Isotopic, 294

- Jacobi identity, 155, 376
 - for the bracket in any ring, 378
- Jacobian matrix, 40
- Jordan Curve Theorem, 21, 435

- Kelley, J., 460, 463, 464
- Kink, 366
- Klein bottle, 18, 435
- Kneser, H., 472
- Kneser, M., 473

- Lang, S., 145
- Laplace's expansion, 230
- Laplacian, 58
- Law of Inertia, Sylvester's, 349
- Leaf, 194
- Leap of faith, 464
- Left invariant
 - form, 394
 - n form, 400
 - vector field, 374
- Left translation, 374
- Length, 243, 305, 312
 - of a curve, 59
- Lie algebra, 376
 - abelian, 376, 382, 395
 - commutative, 376
 - homomorphism of, 380
 - ideal of, 410
 - opposite, 407
- Lie derivative, 150
- Lie group, 371
 - arcwise connected subgroup of, 409
 - closed subgroup of, 391
 - local, 415
 - normal subgroup of, 410
 - topologically isomorphic, 388
- Lie subgroup, 373
- Lie's fundamental theorem
 - first, 414
 - second, 415
 - third, 416
- Limit
 - ordinal, 463
 - set, 60
- Line integral, 239, 243
- Linear differential equations, 165
 - systems of, 171
- Linear transformation
 - adjoint of, 103
 - contraction of, 121
 - positive definite, 104
 - positive semi-definite, 104
- Linking number, 296
- Lipschitz condition, 138
- Littlewood, J. E., 179
- Lives at points, 119
- Lobachevskian non-Euclidean geometry, 368
- Local
 - flow, 144
 - Lie group, 415
 - one-parameter group of local diffeomorphism, 148
 - spanned locally, 179
 - triviality, 71
- Local theory of integral manifolds, 190

- Locally
 - compact, 20
 - connected, 20
 - finite cover, 50
 - isomorphic Lie groups, 382
 - Lipschitz, 139
 - one-one, 13
 - pathwise connected, 20
- Long
 - cylinder, 466
 - line, 465
 - ray, 465
 - closed, 465
 - open, 465
- Lower sum, 283

- MacKenzie, R. E., 106
- Magic, 214
- Manifold, 1, 459
 - analytic, 34
 - atlas for, 28
 - boundary of, 19
 - bounded, 19
 - closed, 19
 - C^r , 34
 - C^0 , 34
 - C^∞ , 29
 - differentiable, 29
 - dimension of, 4
 - imbedding in \mathbb{R}^N , 52
 - integral, 179, 181
 - maximal, 194
 - non-metrizable, 465, 466
 - orientation of, 86
 - smooth, 29
- Manifold-with-boundary, 19
 - C^∞ , 32
- Map
 - between complexes, 421
 - bundle, 73
 - rank of, 40
- Massey, W. S., 3
- Matrix groups, 372
- Maximal integral manifold, 194
- Mayer-Vietoris Sequence, 424
 - for compact supports, 431
- Measure zero, 40, 41
- Mesh, 239
- Metric
 - bi-invariant, 401
 - Euclidean, 305, 315
 - indefinite, 350
 - Riemannian, 308, 311
 - usual, 312
 - spaces, disjoint union of, 4, 20
- Milnor, J. W., 42
- Mod 2 degree, 295
- Möbius strip, 10
 - generalized, 100
- Multi-index, 208
- Multilinear function, 115
- Munkres, J. R., 34, 106

- n -dimensional, 4
- n -forms, left invariant, 400
- n -holed torus, 9
- n -manifold, 4
- n -plane bundle, 71
- n -sphere, 7
- n -torus, 7
- Natural g -valued 1-form, 403
- Natural isomorphism, 108
- Neighborhood, tubular, 345
- Newman, M. H. A., 3
- Nice cover, 438
- Non-bounded, 19
- Non-degenerate, 301
- Non-Euclidean geometry
 - elliptical, 367
 - Lobachevskian, 368
- Non-metrizable manifold, 465, 466
- Non-orientable
 - bundle, 86
 - manifold, 86
- Norm, 303
 - preserving, 304, 372
- Normal
 - bundle, 344

- Normal (*continued*)
 - space, 459
 - subgroup of a Lie group, 410
 - outward unit, 351
- Nowhere zero section, 209
- Odd
 - ordinal, 467
 - relative tensor, 134, 288
 - scalar density, 133, 259
- One-dimensional distribution, 179
- One-dimensional sphere, 6
- One-parameter group
 - of diffeomorphisms, 148
 - of local diffeomorphisms, local, 148
- One-parameter subgroup, 384
- Open
 - long ray, 465
 - map, 60
 - submanifold, 2
- Opposite
 - group, 407
 - Lie algebra, 407
- Order
 - isomorphic, 461
 - isomorphism, 461
 - topology, 465
- Ordered set, 461
- Ordering, 460
- Ordinal numbers, 463
- Orientable
 - bundle, 86
 - manifold, 86
- Orientation
 - of a bundle, 85
 - of a manifold, 86
 - of a vector space, 84
 - preserving, 84, 85, 88, 105, 248
 - reversing, 84, 88, 248
- Orthogonal group, 61, 372
- Orthonormal, 304, 348
- Orthonormalization process, Gram-Schmidt, 304
- Osgood's Theorem, 284
- Outside, 21
- Outward pointing, 260
- Outward unit normal, 351
- Palais, R. S., 100, 225
- Paracompact, 210, 459
 - end, 468
- Parameter curves, special, 167
- Parameterized by arclength, 313
- Partial derivatives, 35
- Partition, 239, 245
 - of unity, 52
- Pathwise connected, 20
- Piecewise smooth, 312
- Pig, yellow, 434
- Poincaré, H., 450
- Poincaré dual, 439
- Poincaré Duality Theorem, 441
- Poincaré-Hopf Theorem, 450
- Poincaré Lemma, 225
- Poincaré upper half-plane, 367
- Point
 - inward, 98
 - outward, 98, 260
- Point-derivation, 39
- Point-finite cover, 60
- Polar coordinates, 36
 - integration in, 266
- Polarization, 304
- Pollack, A., 106
- Positive definite, 104, 301
- Positive element of norm 1, 308
- Positive semi-definite, 104
- Product
 - of vector bundles, 102
 - tensor, 116
- Projection, 7, 30, 32
- Projective
 - plane, 11, 435
 - space, 19, 88
- Proper map, 60, 275
- Prüfer manifold, 467
- Pseudometric, 95

- Quaternions, 100
 - of norm 1, 373
- Radial function, 435
- Rado, T., 472
- Rank
 - of a form, 229
 - of a map, 40, 98
- Rectifiable, 59
- Refinement of a cover, 50
- Regular
 - point, 40
 - space, 459
 - value, 40
- Related vector fields, 190
- Relative
 - scalar, 134, 231
 - tensor, 134, 231, 288
- Reparameterization, 244, 248
- Retraction, 264
 - deformation, 279
- Revolution, surface of, 8, 321
- de Rham, G., 342
- de Rham cohomology vector spaces, 263
 - with compact supports, 268
- de Rham's Theorem, 263, 457
- Riemann
 - integrable, 283
 - integral, 283
 - sum, 283
- Riemannian metric, 308, 311
 - usual, 312
- Right invariant n -form, 400
- Right translation, 374
- Rinow, W., 342
- Roman surface, 17, 26
- Rosenlicht, M., 472
- Rotation group, 62
- Sard's Theorem, 42, 294
- Scalar, relative, 134, 231
- Scalar density, 133
- Schwarz, H., 354
- Schwarz inequality, 303, 362
- Second countable, 459
- Section of a vector bundle, 73
 - zero, 96
- Segment, initial, 462
- Self-adjoint linear transformation, 104
- Semi-definite, positive, 104
- Separate points and closed sets, 95
- Sequence
 - exact, 419, 422
 - of vector bundles, 103
 - Mayer-Vietoris, 424
 - for compact supports, 431
 - of a pair, 433
- Shrinking Lemma, 51
- Shrinking Lemma, 60
- Shuffle permutation, 227
- Simplex
 - of a triangulation, 427
 - singular, 285
- Simply-connected, 287
 - Lie group, 382
- Singular
 - cube, 246
 - simplex, 285
- Skew-symmetric, 201, 378
- Slice, 194
- Slice maps, 54
- Smooth, 28
 - homotopy, 277
 - manifold, 29
 - piecewise, 312
- Smoothly
 - contractible, 220
 - homotopic, 277
 - isotopic, 294
- Solid angle, 290
- Space filling curve, 58
- Spanned locally, 179
- Special linear group, 61
- Special orthogonal group, 62
- Sphere, 7
- Sphere bundle, associated, 451

- Standard
 - n -simplex, 426
 - singular cube, 246
- Star-shaped, 221
- Steiner's surface, 17, 26
- Sternberg, S., 42, 106
- Stokes' Theorem, 253, 261, 285, 352
- Stone-Čech compactification, 468
- Structure constants, 396
- Subalgebra of a Lie algebra, 379
- Subbundle, 198
- Subcover, 50
- Subgroup
 - Lie, 373
 - one-parameter, 384
- Submanifold, 49
 - C^∞ , 49
 - closed, 49
 - immersed, 47
 - open, 2
- Successor ordinal, 464
- Sum of vector bundles, Whitney, 101
- Support, 33, 147
- Surface, 7
 - area, 354
 - integral, 239
 - of revolution, 8, 321
- Sylvester's Law of Inertia, 349
- Symmetric bilinear form, 301
- System of linear differential equations, 171
- σ -compact, 4, 459
- Tangent bundle, 77
- Tangent space of \mathbb{R}^n , 64
- Tangent vector
 - inward pointing, 98
 - of a manifold, 76
 - of \mathbb{R}^n , 64
 - outward pointing, 98, 260
 - to a curve, 63, 66
- Tensor
 - contravariant, 120
 - covariant, 113
 - even relative, 134, 231
 - odd relative, 134, 288
- Tensor field
 - classical definition of, 123
 - contravariant, 120
 - covariant, 113
 - mixed, 121, 122
- Tensor product, 116
- Thom class, 442
- Thom Isomorphism Theorem, 456
- Topological
 - group, 371
 - imbedding, 14
 - immersion, 14, 46
- Topologically isomorphic Lie groups, 388
- Torus, 7, 8
 - n -holed, 7, 9
- Total space, 71
- Totally disconnected, 25
- Transitivity, 460
- Translation
 - left, 374
 - right, 374
- Triangle inequality, 303
- Triangulation, 426
 - simplex of, 427
- Trichotomy, 461
- Trivial vector bundle, 72
- Tubular neighborhood, 345
- Two-holed torus, 8
- Vick, J. W., 3
- Wedge product, 203
- Whitney, H., 106
- Whitney sum, 101



These books were typeset using Donald E. Knuth's \TeX typesetting system, together with Berthold Horn's DVIPSONE PostScript driver. The figures were produced with Adobe Illustrator, and new or modified fonts were created using Fontographer.

The text font is 11 point Monotype Baskerville—though the em-dash has been modified—together with its italic. The elegant swashes of the italic *y* and *f* cause problems in words like *topology* and *apology*, so a special *gy* ligature was added; special *gg* and *gf* ligatures were also required.

Although a Baskerville bold face is unhistorical, bold type was useful in special circumstances—mainly for indicating **defined terms**. The bold face supplied by Monotype, even the “semi-bold”, is obtrusively **extended**, so a non-extended version was created.

The somewhat bold appearance of chapter headings, in 16 point type, results from the linear scaling, as well as the fact that the upper case Baskerville letters are of somewhat heavier weight than the lower case. On the other hand, the tall initial letters beginning each chapter were designed specially, since simple scaling would have made them unpleasantly heavy.

A thicker set of numerals was constructed for use with the upper case lettering in chapter headings and statements of theorems, and special parentheses and other punctuation symbols were also required. Numerous other modifications of this sort, including additional kerns and alterations of set widths, were made for various purposes.

The mathematics fonts are a variation of the *MathTime* fonts, now based on the Monotype Times New Roman family, together with the Monotype Times N R Seven and Times Small Text families—presumably these three families are based on the original designs for Times New Roman, which was created in three essential sizes: 9, 7 and $5\frac{1}{2}$ point.

The italic fonts of these three families were used as the basis for creating the three separate “math italic” fonts—for use at ordinary size, in superscripts, and in second order superscripts. The proportions and weights for these were then used for the three sizes of the symbol font and the other mathematics fonts, including bold symbols, script letters, and additional special symbols, as well as for the extension font and its bold version.

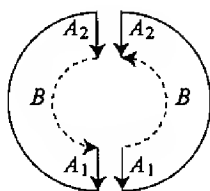
The bold letters for mathematics come directly from the bold fonts of the Times families, while blackboard bold letters were made by hollowing out these bold letters. The Adobe Mathematical Pi 2 font was used for the ordinary sized German Fraktur letters, with suitably modified versions used for superscripts.

The covers, painted by the author in his spare moments, are loosely based on Samuel Taylor Coleridge's poem *The Rime of the Ancient Mariner*.

CORRECTIONS FOR VOLUME I

pg. 3, line 3—: change $\bar{d}_i(x, y) < 1$ to $\bar{d}_i(x, y) \leq 1$.

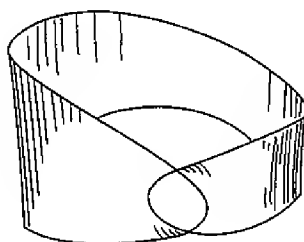
pg. 14: relabel the lower left part of the central figure as



pg. 19: replace the next-to-last paragraph with the following: .

The set of points in a manifold-with-boundary that do not have a neighborhood homeomorphic to \mathbb{R}^n (but only one homeomorphic to \mathbb{H}^n) is called the **boundary** of M and is denoted by ∂M . Equivalently, $x \in \partial M$ if and only if there is a neighborhood V of x and a homeomorphism $\phi : V \rightarrow \mathbb{H}^n$ such that $\phi(x) = 0$. If M is actually a manifold, then $\partial M = \emptyset$, and ∂M itself is always a manifold (without boundary).

pg. 22: Replace the top left figure with



pg. 43: Replace the last line and displayed equation with the following:

Since rank $f = k$ in a neighborhood of p , the lower rectangle in the matrix

$$\left(\frac{\partial(v^i \circ f)}{\partial x^j} \right) = \begin{pmatrix} \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} & 0 \\ \times & \begin{matrix} D_{k+1}\psi^{k+1} \dots D_{k+1}\psi^m \\ \vdots \\ D_n\psi^{k+1} \dots D_n\psi^m \end{matrix} \end{pmatrix}$$

pg. 60, Problem 30: Change part (f) and add part (g):

(f) If M is a connected manifold, there is a proper map $f : M \rightarrow \mathbb{R}$; the function f can be made C^∞ if M is a C^∞ manifold.

(g) The same is true if M has at most countably many components.

pg. 61, Problem 32: For clarity, restate part (c) as follows:

(c) This is false if $f : M_1 \rightarrow \mathbb{R}$ is replaced with $f : M_1 \rightarrow N$ for a disconnected manifold N .

pg. 70: Replace the last two lines of page 70 and the first two lines of page 71 with the following:

theorem of topology). If there were a way to map $T(M, i)$, fibre by fibre, homeomorphically onto $M \times \mathbb{R}^2$, then each v_p would correspond to $(p, v(p))$ for some $v(p) \in \mathbb{R}^2$, and we could continuously pick $w(p) \in \mathbb{R}^2$, corresponding to a dashed vector, by using the criterion that $w(p)$ should make a positive angle with $v(p)$.

pg. 78: the third display should read:

$$0 = \ell(0) = \ell(fh) = f(p)\ell(h) + h(p)\ell(f) = 0 + \ell(f).$$

pg. 103, Problem 29(d). Add the hypothesis that M is orientable.

pg. 117: After the next to last display, $\bar{A}(X_1, \dots, X_k)(p) = A(p)(X_1(p), \dots, X_k(p))$, add:

If A is C^∞ , then \bar{A} is C^∞ , in the sense that $\bar{A}(X_1, \dots, X_k)$ is a C^∞ function for all C^∞ vector fields X_1, \dots, X_k .

pg. 118: Add the following to the statement of the theorem: If \mathcal{A} is C^∞ , then A is also.

pg. 119: Add the following at the end of the proof:

Smoothness of A follows from the fact that the function $A_{i_1 \dots i_k}$ is $\mathcal{A}(\partial/\partial x_{i_1}, \dots, \partial/\partial x_{i_k})$.

pg. 131, Problem 9: Let F be a covariant functor from \mathbf{V} ,

pg. 133. Though there is considerable variation in terminology, what are here called “odd scalar densities” should probably simply be called “scalar densities”; what are called “even scalar densities” might best be called “signed scalar densities”.

In part (c) of Problem 10, we should be considering the h of part (a), not the h of part (b)! Thus conclude that the bundle of signed scalar densities (*not* the scalar densities) is not trivial if M is not orientable.

pg. 134. Extending the changed terminology from pg. 133, we should probably speak of the bundle of “signed tensor densities of type $\binom{k}{l}$ and weight w ” (though sometimes the term relative tensor is used instead, restricting densities to those of weight 1), when the transformation rule involves $(\det A)^w$, omitting the modifier “signed” when it involves $|\det A|^w$.

pg. 143. The hypothesis of Theorem 3 should be changed so that it reads:

Let $x \in U$ and let α_1, α_2 be two maps on some open interval I such that $\alpha_1(I), \alpha_2(I) \subset U$,

$$\alpha_i'(t) = f(\alpha_i(t)) \quad i = 1, 2$$

and

$$\alpha_1(t_0) = \alpha_2(t_0) \quad \text{for some } t_0 \in I.$$

And the first sentence of the proof should be deleted.

pg. 177. Problem 17, part (d) should begin:

(d) Let $f: M \rightarrow N$, and suppose that $f_{*p} = 0$. For $X_p, Y_p \in M_p$ and

pg. 198. In Problem 5, we must also assume that each $\Delta_i \oplus \Delta_j$ is integrable.

pg. 226. In the comutative diagram, the lower right entry should be “ I -forms on N ”.

pg. 233. The reference “pg. V.375” refers to pg. 375 of Volume V.

pg. 237. In Problem 26, replace parts (b) and (c) with:

(b) Determine the i^{th} component of $v_1 \times \dots \times v_{n-1}$ in terms of the $(n-1) \times (n-1)$ submatrices of the matrix

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

In particular, for \mathbb{R}^3 , show that

$$v \times w = (v^2 w^3 - v^3 w^2, v^3 w^1 - v^1 w^3, v^1 w^2 - v^2 w^1).$$

pg. 292. In Problem 20, the condition $U_i \cap U_j \neq \emptyset$ should be $U_i \cap U_{i+1} \neq \emptyset$.

pg. 408. Problem 16 (b) should read: “For any Lie group G , show that . . .”.

pp. 408–410. For consistency with standard usage, Aut should be replaced with Aut , and then replace *End* with *End*. In part (g) of Problem 19, add the hypothesis that H is a connected Lie subgroup.

pg. 411. The display in Problem 21, part (c) should read:

$$(-1)^{km}[\omega \wedge [\eta \wedge \lambda]] + (-1)^{ki}[\eta \wedge [\lambda \wedge \omega]] + (-1)^{lm}[\lambda \wedge [\omega \wedge \eta]] = 0.$$

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME TWO
Third Edition



MICHAEL SPIVAK

PUBLISH OR PERISH, INC.



Houston, Texas 1999

Publish or Perish, Inc.
www.mathpop.com

Copyright © 1970, 1979, 1999 by Michael Spivak
All Rights Reserved

Volume 1 ISBN 0-914098-70-5
Volume 2 ISBN 0-914098-71-3
Volume 3 ISBN 0-914098-72-1
Volume 4 ISBN 0-914098-73-X
Volume 5 ISBN 0-914098-74-8

Printed in the United States of America

For
Harry
Josh
and
Marc

PREFACE

Though this volume begins the study of modern differential geometry in earnest, it ends just when we have gotten to the fanciest definition of a connection, but have hardly begun to start spouting theorems. A glance at the Table of Contents will show that the semi-historical path promised in Volume I really has been followed. The most decisive encounters with classical differential geometry occur in Chapters 3 and 4, which present the classical papers and then explain them. While it is possible to get through this volume without reading any of the classical works themselves, the easy way out certainly misses all the fun!

There are no Problem sets in this volume, which is a shame, but much of the material doesn't lend itself to problems, and even if it did, I would have gone berserk trying to produce them in any reasonable length of time. As compensation for the lack of Problems, the final volume contains a comprehensive bibliography of the literature of Differential Geometry, including texts where problems may be found.

TABLE OF CONTENTS

Although the chapters are not divided into sections,
the listing for each chapter gives some indication
which topics are treated, and on what pages.

CHAPTER 1. CURVES IN THE PLANE AND IN SPACE

Curvature of plane curves	1
Convex curves	12
Curvature and torsion of space curves	24
The Serret-Frenet formulas	34
The natural form on a Lie group	36
Classification of plane curves under the group of special affine motions	39
Classification of curves in \mathbb{R}^n	44

CHAPTER 2. WHAT THEY KNEW ABOUT SURFACES BEFORE GAUSS

Euler's Theorem	50
Meusnier's Theorem	52

CHAPTER 3. THE CURVATURE OF SURFACES IN SPACE

A. HOW TO READ GAUSS	55
B. GAUSS' THEORY OF SURFACES	
The Gauss map	112
Gaussian curvature	114
The Weingarten map; the first and second fundamental forms . . .	122
The Theorema Egregium	129
Geodesics on a surface	134
The metric in geodesic polar coordinates	136
The integral of the curvature over a geodesic triangle	141

Addendum. The formula of Bertrand and Puiseux; Diquet's formula	145
--	-----

CHAPTER 4. THE CURVATURE OF HIGHER DIMENSIONAL MANIFOLDS

A. AN INAUGURAL LECTURE	149
“On the Hypotheses which lie at the Foundations of Geometry” . . .	151
B. WHAT DID RIEMANN SAY?	163
The form of the metric in Riemannian normal coordinates	166
C. A PRIZE ESSAY	181
D. THE BIRTH OF THE RIEMANN CURVATURE TENSOR	
Necessary conditions for a metric to be flat	184
The Riemann curvature tensor	189
Sectional Curvature	194
The Test Case; first version	197
Addendum. Finsler metrics	200

CHAPTER 5. THE ABSOLUTE DIFFERENTIAL CALCULUS (THE RICCI CALCULUS); OR, THE DEBAUCH OF INDICES

Covariant derivatives	209
Ricci's Lemma	213
Ricci's identities	214
The curvature tensor	215
The Test Case; second version	217
Classical connections	221
The torsion tensor	221
Geodesics	223
Bianchi's identities	224

CHAPTER 6. THE ∇ OPERATOR

Kozul connections	227
Covariant derivatives	229
Parallel translation	234
The torsion tensor	236
The Levi-Civita connection	238

The curvature tensor	239
The Test Case; third version	241
Bianchi's identities	244
Geodesics	246
The First Variation Formula	247
Addendum 1. Connections with the same geodesics	249
Addendum 2. Riemann's invariant definition of the curvature tensor	254

CHAPTER 7. THE REPÈRE MOBILE (THE MOVING FRAME)

Moving frames	259
The structural equations of Euclidean space	261
The structural equations of a Riemannian manifold	267
The Test Case; fourth version	268
Adapted frames	270
The structural equations in polar coordinates	272
The Test Case; fifth version	274
The Test Case; sixth version	275
"The curvature determines the metric"	277
The 2-dimensional case	279
Cartan connections	281
Covariant derivatives and the torsion and curvature tensors	285
Bianchi's identities	288
Addendum 1. Manifolds of constant curvature	290
Schur's Theorem	291
The form of the metric in normal coordinates	295
Addendum 2. Conformally equivalent manifolds	296
Addendum 3. É. Cartan's treatment of normal coordinates	302

CHAPTER 8. CONNECTIONS IN PRINCIPAL BUNDLES

Principal bundles	305
Lie groups acting on manifolds	309
A new definition of Cartan connections	311
Ehresmann connections	315
Lifts	317
Parallel translation and covariant derivatives	319

The covariant differential and the curvature form	324
The dual form and the torsion form	324
The structural equations	327
The torsion and curvature tensors	329
The Test Case; seventh version	333
Bianchi's identities	334
Summary	337
Addendum 1. The tangent bundle of $F(M)$	342
Addendum 2. Complete connections	344
Addendum 3. Connections in vector bundles	346
Addendum 4. Flat connections	349
NOTATION INDEX	351
INDEX	355

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME TWO

CHAPTER 1

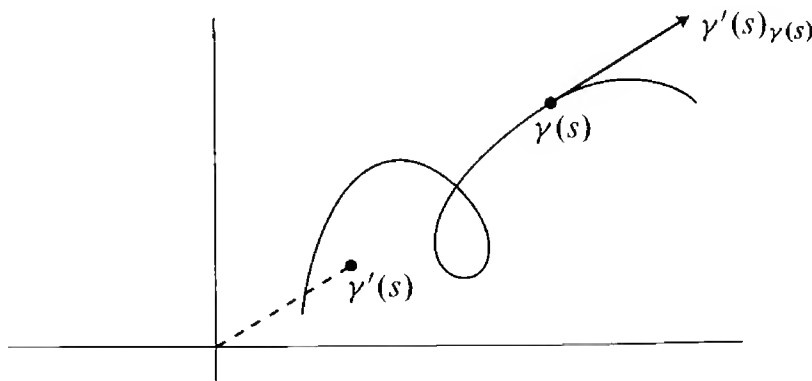
CURVES IN THE PLANE AND IN SPACE

Volume I of these notes represents the “differential” part of differential geometry. In this volume we finally get down to some geometry. For the present we are going to study only the simplest geometric objects, curves, and at first our approach will be terribly geometric. Nevertheless, the results of this chapter span a couple hundred years, and we will end with some very modern looking constructions.

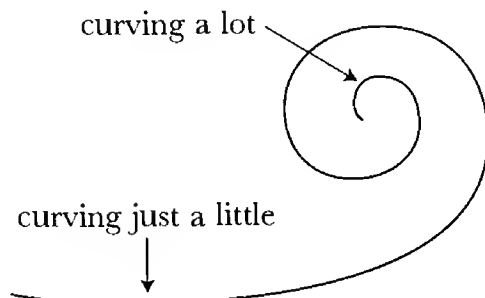
We begin by considering only curves in the plane, and we further restrict our attention to curves $c: [a, b] \rightarrow \mathbb{R}^2$ which are *immersions*, i.e., which satisfy $c'(t) = dc/dt \neq 0$ for all $t \in [a, b]$. For these curves, the arclength function $s: [a, b] \rightarrow \mathbb{R}$,

$$s(t) = \int_a^t |c'(u)| du,$$

is a diffeomorphism $s: [a, b] \rightarrow [0, L]$, where $L = \text{length of } c$. The curve $\gamma = c \circ s^{-1}$ is then a reparameterization of c ; clearly γ is parameterized by arclength, $|\gamma'(s)| = 1$. We have just introduced a convention to be used throughout the chapter: for curves $\gamma: [a, b] \rightarrow \mathbb{R}^2$ with unit tangent vectors, we will usually denote a typical point in $[a, b]$ by s , even at the risk of confusing it with the arclength function defined above. We also emphasize that throughout this chapter $\gamma'(s)$ just denotes a vector in \mathbb{R}^3 , not a tangent vector in $\mathbb{R}^3_{\gamma(s)}$, even though we will often draw it that way.

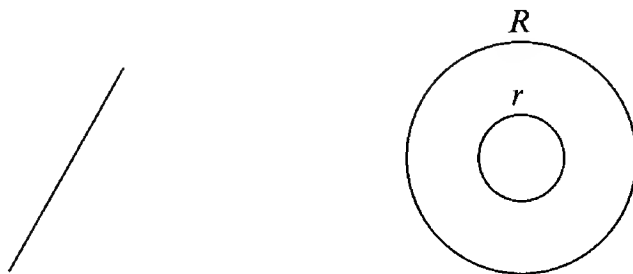


For an immersed curve $c: [a, b] \rightarrow \mathbb{R}^2$, we would like a way of measuring the amount that c is “curving” at any point. No matter how vague this, as yet



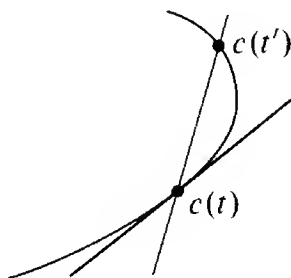
intuitive, term may be, we will surely all agree that

- (a) a straight line is not curving at all,
- (b) a circle of radius $R > r$ is curving less than the circle of radius r .

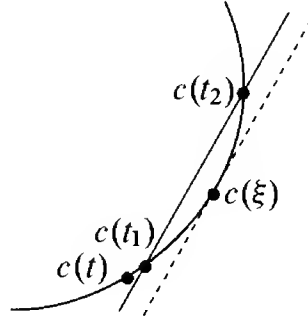


If we have to attach a numerical measure of curvature to these particular curves, it seems reasonable to define the curvature of a straight line at any point to be 0, and to define the curvature of a circle of radius r to be $1/r$ at any point.

From these special cases we want to develop a definition that works for any curve. We take as a clue the procedure which we use in a similar, but simpler, case. We can easily define what we mean by the *direction* of a curve $c: [a, b] \rightarrow \mathbb{R}^2$ at any time t . This direction is determined by the tangent line at $c(t)$, which is the limit of lines through $c(t)$ and $c(t')$ as $t' \rightarrow t$. This limit exists if $c'(t)$



exists and is non-zero. If, moreover, c' is continuous at t , then this limit line can be described as the limit, as $t_1, t_2 \rightarrow t$, of the line through $c(t_1)$ and $c(t_2)$; for this line is parallel to the tangent through $c(\xi)$ for some ξ between t_1 and t_2 .



In order to determine the curvature of a curve we follow an analogous procedure. We find the circle which passes through three points $c(t_1), c(t_2), c(t_3)$ and then see if this circle approaches a limiting circle as $t_1, t_2, t_3 \rightarrow t$. If it does, we can define the curvature of c at t to be the reciprocal of the radius of this circle.

Before proving a precise theorem, we simply try to determine the position of this circle, assuming it does exist. This can be done as follows. First of all, since c is an immersion, it is locally one-one, so for distinct t_1, t_2, t_3 near t , the points $c(t_1), c(t_2), c(t_3)$ are distinct. To be specific, let us say that $t_1 < t_2 < t_3$. Suppose also that $c(t_1), c(t_2), c(t_3)$ do not lie on a straight line, so that there is a unique circle through these three points, with center $C(t_1, t_2, t_3)$. Consider the function

$$t \mapsto \langle c(t) - C(t_1, t_2, t_3), c(t) - C(t_1, t_2, t_3) \rangle.$$

At t_1, t_2, t_3 this function has the same value, the square of the radius of the circle through $c(t_1), c(t_2), c(t_3)$. So its derivative must be 0 at points $\xi_1 \in (t_1, t_2)$ and $\xi_2 \in (t_2, t_3)$:

$$(1) \quad 0 = \langle c'(\xi_i), c(\xi_i) - C(t_1, t_2, t_3) \rangle, \quad \xi_i \in (t_i, t_{i+1}), \quad i = 1, 2.$$

Similarly, the function $t \mapsto \langle c'(t), c(t) - C(t_1, t_2, t_3) \rangle$ must have derivative 0 at some point $\eta \in (\xi_1, \xi_2)$:

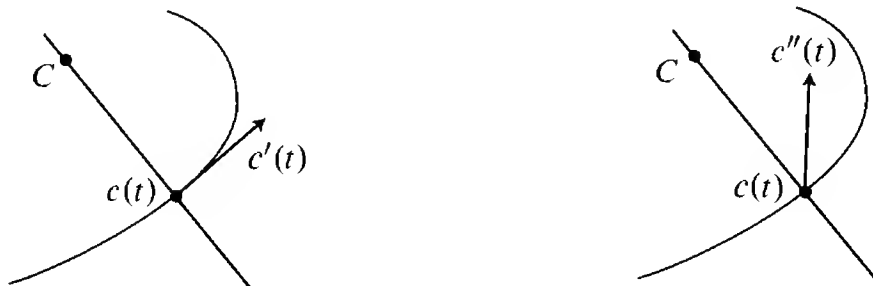
$$(2) \quad \langle c''(\eta), c(\eta) - C(t_1, t_2, t_3) \rangle = -\langle c'(\eta), c'(\eta) \rangle.$$

Now, if the points $C(t_1, t_2, t_3)$ approach a point C as $t_1, t_2, t_3 \rightarrow t$, and if c'' is continuous, then (1) and (2) clearly imply that

$$(1') \quad \langle c'(t), c(t) - C \rangle = 0$$

$$(2') \quad \langle c''(t), c(t) - C \rangle = -\langle c'(t), c'(t) \rangle.$$

The first of these equations shows that the circle through $c(t)$ with center C must be tangent to c at $c(t)$, which is certainly to be expected. Thus C is already



restricted to lie along a certain line. If $c''(t)$ is not a multiple of $c'(t)$, the second equation then determines C , since it tells us the inner product of $c(t) - C$ with $c''(t)$. If $c''(t)$ is a multiple of $c'(t)$, then we obtain the contradiction

$$0 \neq -\langle c'(t), c'(t) \rangle = \langle c''(t), c(t) - C \rangle = \text{constant} \cdot \langle c'(t), c(t) - C \rangle = 0.$$

In other words, if $c''(t)$ is a multiple of $c'(t)$, this limiting position cannot exist.

Although equations (1'), (2') could be solved for C , we can make things a lot easier for ourselves by considering a curve c parameterized by arclength, so that $|c'(s)| = 1$. (This means that $c'(s)$ always lies on the unit circle $S^1 \subset \mathbb{R}^2$, even though we often picture it instead as a tangent vector.) Now the equation

$$\langle c'(s), c'(s) \rangle = 1$$

can be differentiated to give

$$(*) \quad \langle c''(s), c'(s) \rangle = 0.$$

In other words, $c''(s)$ is always perpendicular to $c'(s)$. In particular, $c''(s)$ is a multiple of $c'(s)$ only when $c''(s) = 0$. Equations (1') and (*) show that $c''(s)$ and $c(s) - C$ are both perpendicular to $c'(s)$. If $c''(s) \neq 0$, then we can write $c(s) - C = a \cdot c''(s)$. Substituting in (2') gives

$$a \cdot \langle c''(s), c''(s) \rangle = -\langle c'(s), c'(s) \rangle = -1.$$

Since we also have

$$|c(s) - C| = |a| \cdot |c''(s)|,$$

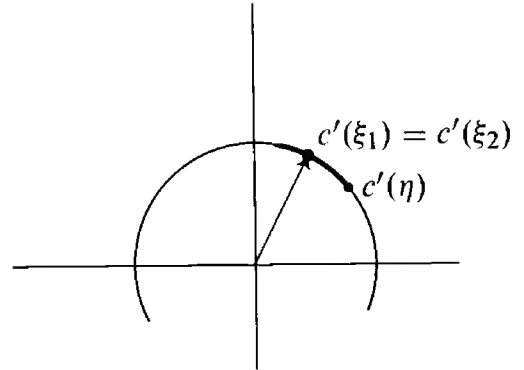
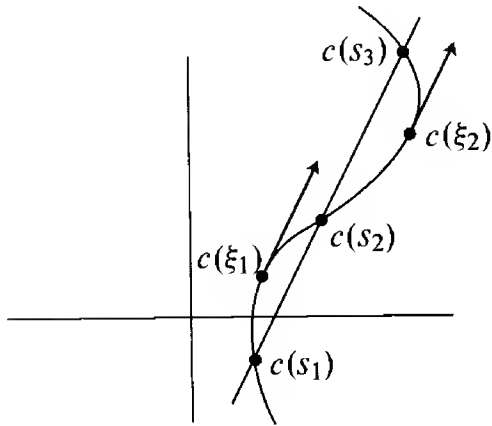
we easily deduce that

$$|c(s) - C| = \frac{1}{|c''(s)|}.$$

In other words, our circle is perpendicular to c at $c(s)$, and has radius $1/|c''(s)|$. Its curvature, and hence the curvature of c at s , is thus $|c''(s)|$. We are ready to reverse the order of this reasoning, and take care of details which we have ignored.

1. **THEOREM.** Let $c: [a, b] \rightarrow \mathbb{R}^3$ be a C^2 curve parameterized by arclength. If $c''(s) \neq 0$, then for s_1, s_2, s_3 sufficiently close to s , the points $c(s_1), c(s_2), c(s_3)$ do not lie on a straight line. As $s_1, s_2, s_3 \rightarrow s$ the unique circle through the points $c(s_i)$ approaches a circle passing through $c(s)$, whose radius is $1/|c''(s)|$, and whose center lies on the line through $c(s)$ perpendicular to the tangent line through $c(s)$. If $c''(s) = 0$, then, even if the points $c(s_i)$ do not lie on a line, the circles through them do not approach a limiting circle.

PROOF. We first show that if $c''(s) \neq 0$, then the points $c(s_1), c(s_2), c(s_3)$ cannot lie on a line for s_1, s_2, s_3 arbitrarily close to s . Whenever the points $c(s_i)$ lie on a line, for $s_1 < s_2 < s_3$, there are points $\xi_1 \in (s_1, s_2)$ and $\xi_2 \in (s_2, s_3)$ where the tangent lines are parallel to this line (Cauchy mean value theorem).



This means that the curve c' in S^1 has $c'(\xi_1) = c'(\xi_2)$. For ξ_1 and ξ_2 close enough to s , the image $c'([\xi_1, \xi_2])$ can't be the whole of S^1 . So there is a point $\eta \in (\xi_1, \xi_2)$ where $c'(\eta)$ is furthest from $c'(\xi_i)$ in some direction along the circle. It follows that $c''(\eta) = 0$. This cannot happen for η arbitrarily close to s , so the points $c(s_1), c(s_2), c(s_3)$ cannot lie on a line for s_1, s_2, s_3 arbitrarily close to s .

Now let C be the unique point satisfying

$$(*) \quad \begin{aligned} \langle c'(s), c(s) - C \rangle &= 0 \\ \langle c''(s), c(s) - C \rangle &= -\langle c'(s), c'(s) \rangle = -1. \end{aligned}$$

For $s_1 < s_2 < s_3$ close to s , let $C(s_1, s_2, s_3)$ be the center of the unique circle through the points $c(s_i)$. We have already seen that

$$\begin{aligned} \langle c'(\xi), c(\xi) - C(s_1, s_2, s_3) \rangle &= 0 & \xi \in (s_1, s_3) \\ \langle c''(\eta), c(\eta) - C(s_1, s_2, s_3) \rangle &= -\langle c'(\eta), c'(\eta) \rangle = -1 & \eta \in (s_1, s_3). \end{aligned}$$

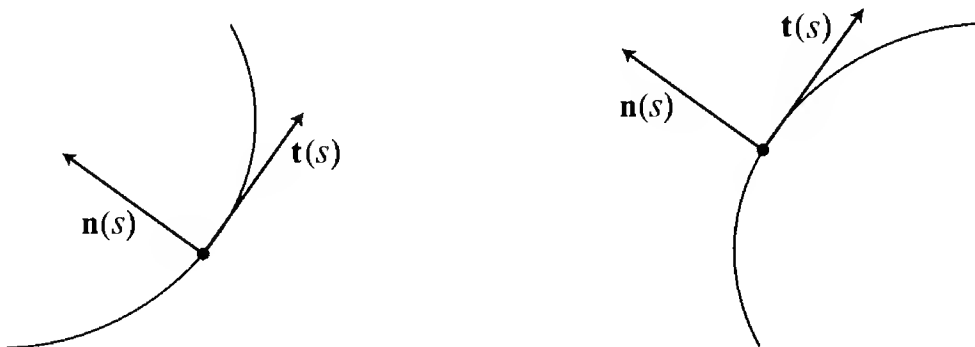
Since $c'(\xi) \rightarrow c'(s)$ and $c''(\eta) \rightarrow c''(s)$ as $s_i \rightarrow s$, comparison of these equations with $(*)$ shows that $C(s_1, s_2, s_3)$ must approach C .

We have already shown that if $c''(s) = 0$, then $C(s_1, s_2, s_3)$ cannot approach a limiting position as $s_i \rightarrow s$. ♦

The circle determined in Theorem 1 is called the **osculating circle** of c at s (“osculate” means to kiss). It is clearly the circle which best approximates the



curve c at $c(s)$. This suggests that we define $|c''(s)|$ to be the curvature of c at s ; even for $c''(s) = 0$ this gives the result we would like. With this definition, curvature would always be non-negative, but we can modify the definition slightly so that we obtain a signed curvature. We will henceforth use the notation $\mathbf{t}(s)$ for $c'(s)$, the unit tangent vector of c at s . We use this notation only for curves parameterized by arclength; recall once again that $\mathbf{t}(s) \in S^1$, even though we usually draw it as an element of $\mathbb{R}^3_{c(s)}$. We then define $\mathbf{n}(s)$, the unit normal at s , to be the unit vector such that $\mathbf{n}(s)$ is perpendicular to $\mathbf{t}(s)$, and $[\mathbf{t}(s), \mathbf{n}(s)]$ is the standard orientation of \mathbb{R}^2 . Thus $\mathbf{n}(s) = (-\mathbf{t}^2(s), \mathbf{t}^1(s))$. Now we define



the **curvature** $\kappa(s)$ of c at s by the equation

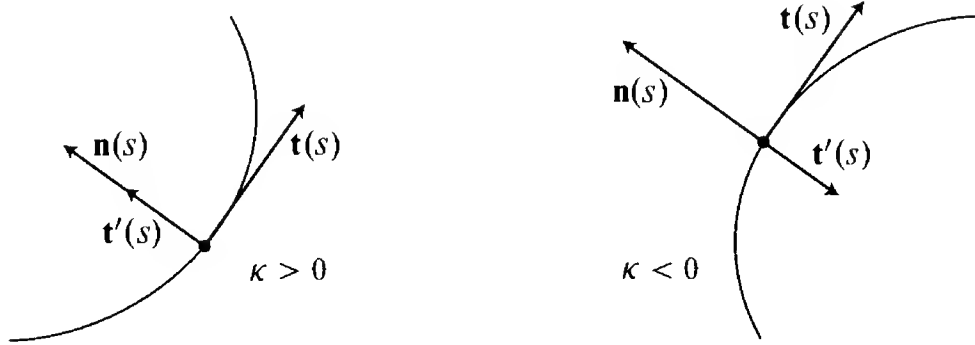
$$\mathbf{t}'(s) = \kappa(s) \cdot \mathbf{n}(s).$$

Notice that

$$|\kappa(s)| = |\mathbf{t}'(s)| = |c''(s)|,$$

so that $1/|\kappa(s)|$ is the radius of the osculating circle at s , for $\kappa(s) \neq 0$. The

significance of the sign of κ is indicated in the figure below.



For any curve c we can define a curve \bar{c} “going in the opposite direction” by $\bar{c}(s) = c(-s)$. If we use $\bar{\mathbf{t}}(s)$ for $\bar{c}'(s)$, then clearly

$$\begin{aligned}\bar{\mathbf{t}}(s) &= -\mathbf{t}(-s), & \text{hence } \bar{\mathbf{n}}(s) &= -\mathbf{n}(-s) \\ \bar{\mathbf{t}}'(s) &= \mathbf{t}'(-s).\end{aligned}$$

This shows that the curvature $\bar{\kappa}(s)$ of \bar{c} at s is

$$\bar{\kappa}(s) = -\kappa(-s).$$

One can see this change of sign in the figure above. Traversing the left curve in the opposite direction, and turning the figure upside down, one obtains the curve on the right.

This relation can also be seen from explicit formulas for κ , which are sometimes useful to have. To write these, we abandon our usual practice of indicating component functions with superscripts, and instead write $c(s) = (c_1(s), c_2(s))$. We have, of course, $|\kappa(s)| = \sqrt{(c_1''(s))^2 + (c_2''(s))^2}$, but we can also develop a formula for $\kappa(s)$ itself. Since $|\kappa(s)|$ is the length of $c''(s)$, which is perpendicular to the unit vector $c'(s)$, clearly $|\kappa(s)|$ is also the area of the rectangle spanned by $c'(s)$ and $c''(s)$. This area is given by $\det(c'(s), c''(s))$ which, moreover,



clearly has the same sign as $\kappa(s)$. So

$$\kappa(s) = \det \begin{pmatrix} c_1'(s) & c_1''(s) \\ c_2'(s) & c_2''(s) \end{pmatrix} = [c_1'c_2'' - c_2'c_1''](s).$$

Until now we have been working exclusively with curves parameterized by arclength. Theoretically this is sufficient, since we consider only curves which can be reparameterized by arclength. In practice, however, it is usually very inconvenient to actually perform this reparameterization, which involves the inverse of a function defined as an integral. Moreover, since our formula involves only derivatives, only the integrand of this integral should play any crucial role. Consider any (immersed) curve $c : [a, b] \rightarrow \mathbb{R}^2$. Letting $s : [a, b] \rightarrow [0, L]$ be arclength, and defining $\gamma = c \circ s^{-1}$, we have

$$\begin{aligned} c &= \gamma \circ s \\ \frac{dc}{dt} &= \gamma'(s) \frac{ds}{dt} \quad \left(\text{i.e., } \frac{dc}{dt}(t) = \gamma'(s(t)) \frac{ds}{dt}(t) \right) \\ \frac{d^2c}{dt^2} &= \gamma''(s) \left(\frac{ds}{dt} \right)^2 + \gamma'(s) \frac{d^2s}{dt^2}. \end{aligned}$$

Thus

$$\begin{aligned} \gamma'(s) &= \frac{dc}{dt} \bigg/ \frac{ds}{dt} \\ \gamma''(s) &= \frac{\frac{d^2c}{dt^2} - \frac{d^2s}{dt^2} \frac{dc}{dt} \bigg/ \frac{ds}{dt}}{\left(\frac{ds}{dt} \right)^2} = \frac{\frac{d^2c}{dt^2} - \alpha \frac{dc}{dt}}{\left(\frac{ds}{dt} \right)^2}, \quad \text{say.} \end{aligned}$$

Denoting dc_i/dt by $\dot{c}_i(t)$, we have for the curvature $\kappa : [a, b] \rightarrow \mathbb{R}^2$ of c ,

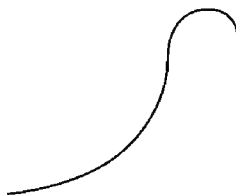
$$\begin{aligned} \kappa(t) &= \text{curvature of } \gamma \text{ at } s = s(t) \\ &= \det(\gamma'(s), \gamma''(s)) \\ &= (ds/dt)^{-3} \det(\dot{c}(t), \ddot{c}(t) - \alpha \dot{c}(t)) \\ &= (ds/dt)^{-3} \det(\dot{c}(t), \ddot{c}(t)). \end{aligned}$$

Thus

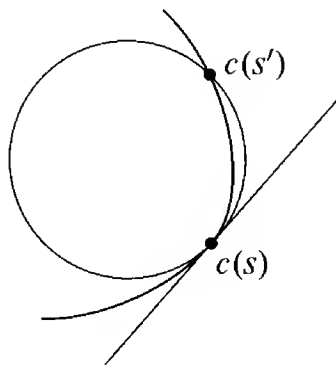
$$\boxed{\kappa = \frac{\dot{c}_1 \ddot{c}_2 - \dot{c}_2 \ddot{c}_1}{(\dot{c}_1^2 + \dot{c}_2^2)^{3/2}}.}$$

Naturally, this formula becomes meaningless for a non-immersed curve, where $\dot{c}_1(t) = \dot{c}_2(t) = 0$. In such cases, we should not generally expect to

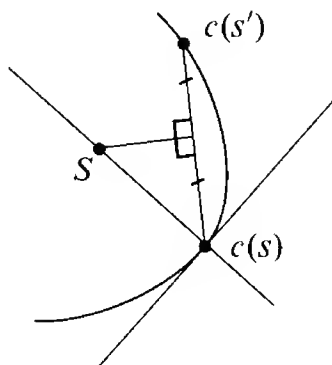
have a meaningful notion of curvature, for there is nothing to stop the curve from “changing its curvature” at this point.



Before examining the significance of the curvature function κ further, we pause for an observation. Since the osculating circle is the limiting position of the circle through $c(s_1), c(s_2), c(s_3)$ *no matter how* $s_1, s_2, s_3 \rightarrow s$, it is clearly also the limiting position as $s' \rightarrow s$ of the circle which is tangent to c at $c(s)$ and which passes through $c(s')$. To see this, we just choose s_1, s_2 much closer



to s than s' is. The center of the osculating circle can thus be described as the limiting position of the point S in the figure below as $s' \rightarrow s$. These descriptions of curvature go back to Huygens, Leibniz, and Newton.



We began our discussion of curvature by considering straight lines and circles, which were our original models of curves which ought to have constant curvature. It is certainly clear that our definitions do assign curvature 0 to a

straight line c , for $c'' = 0$ if c is parameterized by arclength; and absolute curvature $|\kappa| = R$ to a circle of radius R , since the circle is its own osculating circle. A really satisfactory measure of curvature ought to assign constant curvature to these curves alone. It is clear that if c , parameterized by arclength, has curvature 0 everywhere, so that $c'' = 0$, then c' is constant, so c is a straight line. The analysis of a curve c with non-zero constant curvature κ (which we might as well assume positive) becomes frustratingly complicated if approached in too straightforward a way. We assume, as usual, that c is parameterized by arclength. Let us introduce the components of the unit tangent vector curve,

$$\mathbf{t}(s) = (\alpha(s), \beta(s)).$$

Then

$$\alpha'^2 + \beta'^2 = \kappa^2,$$

so

$$(1) \quad \alpha' \alpha'' + \beta' \beta'' = 0.$$

Recall also that $\langle \mathbf{t}(s), \mathbf{t}'(s) \rangle = \langle c'(s), c''(s) \rangle = 0$, so

$$(2) \quad \alpha \alpha' + \beta \beta' = 0.$$

Equations (1) and (2) show that (α, β) and (α'', β'') are always perpendicular to (α', β') , so one is a multiple of the other,

$$(3) \quad \begin{aligned} \alpha''(s) &= \mu(s)\alpha(s) \\ \beta''(s) &= \mu(s)\beta(s). \end{aligned}$$

Moreover, differentiating (2) gives

$$\alpha'^2 + \beta'^2 + \alpha \alpha'' + \beta \beta'' = 0$$

or

$$(4) \quad \kappa^2 + \alpha \alpha'' + \beta \beta'' = 0.$$

Substituting from (3), and using $\alpha^2 + \beta^2 = 1$, gives $\mu(s) = -\kappa^2$. Thus,

$$(5) \quad \begin{aligned} \alpha'' + \kappa^2 \alpha &= 0 \\ \beta'' + \kappa^2 \beta &= 0. \end{aligned}$$

The solutions of the differential equation in (5) are $s \mapsto a \sin \kappa s + b \cos \kappa s$, which can also be written as $s \mapsto A \sin(\kappa s + B)$. In order to have $\alpha^2 + \beta^2 = 1$, we clearly need $A = 1$ and

$$\begin{aligned}\alpha(s) &= \sin(\kappa s + B) \\ \beta(s) &= \cos(\kappa s + B).\end{aligned}$$

Thus \mathbf{t} traverses a circle, which implies that c itself is a circle of radius $1/\kappa$.

The complicated calculations in the preceding paragraph conceal a basic principle which is much simpler. Suppose that we are given an arbitrary continuous function $\kappa: [a, b] \rightarrow \mathbb{R}$. We can ask how many arclength parameterized curves $c: [a, b] \rightarrow \mathbb{R}^2$ there are with curvature function equal to κ , without necessarily trying to find a specific formula for these curves. If there is one such curve, then there are automatically others, for translating or rotating a curve will not change its curvature. However, this is the only extent to which the curve is not determined.

2. THEOREM. Let $\kappa: [a, b] \rightarrow \mathbb{R}$ be continuous. Then there is a curve $c: [a, b] \rightarrow \mathbb{R}^2$, parameterized by arclength, whose curvature at s is $\kappa(s)$ for all $s \in [a, b]$. Moreover, if c and \bar{c} are two such curves, then $\bar{c} = A \circ c$ where A is some proper Euclidean motion (a translation followed by a rotation [an element of $\text{SO}(2)$]).

PROOF. Theorem I.5-17 implies that there is a function $\mathbf{t}: [a, b] \rightarrow \mathbb{R}^2$ with

$$(*) \quad \mathbf{t}'(s) = \kappa(s) \cdot (-\mathbf{t}_2(s), \mathbf{t}_1(s)).$$

We can choose $\mathbf{t}(a)$ arbitrarily; choose it to be a unit vector. Now

$$\begin{aligned}(\mathbf{t}_1^2 + \mathbf{t}_2^2)' &= 2\mathbf{t}_1\mathbf{t}_1' + 2\mathbf{t}_2\mathbf{t}_2' \\ &= 2\langle (\mathbf{t}_1, \mathbf{t}_2), (\mathbf{t}_1', \mathbf{t}_2') \rangle \\ &= 2\langle (\mathbf{t}_1, \mathbf{t}_2), \kappa(-\mathbf{t}_2, \mathbf{t}_1) \rangle \\ &= 0.\end{aligned}$$

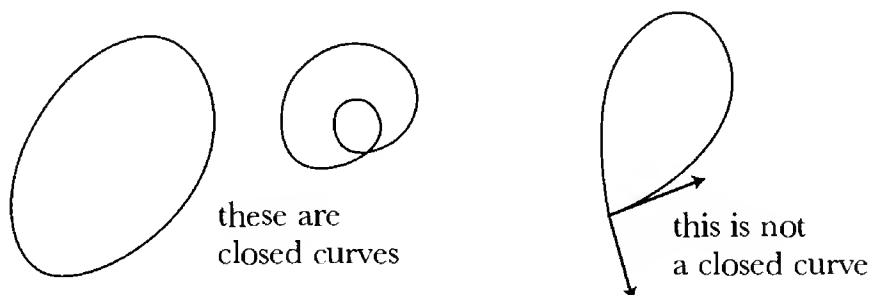
So $\mathbf{t}(s)$ is a unit vector for all s . There is, again by Theorem I.5-17, a curve $c: [a, b] \rightarrow \mathbb{R}^2$ with $c'(s) = \mathbf{t}(s)$. Since $\mathbf{t}(s)$ is always a unit vector, c is parameterized by arclength. Equation $(*)$ then says that $\mathbf{t}'(s) = \kappa(s) \cdot \mathbf{n}(s)$, so that $\kappa(s)$ is the curvature of c at s .

If c and \bar{c} have the same curvature functions κ , then their unit tangent vectors \mathbf{t} and $\bar{\mathbf{t}}$ both satisfy $(*)$. Now if \mathbf{t} is any solution of $(*)$, clearly $B \circ \mathbf{t}$ is also, for any rotation B . Choosing B so that $B(\mathbf{t}(a)) = \bar{\mathbf{t}}(a)$, and using the uniqueness of solutions of $(*)$ with a given initial condition, we see that $\bar{\mathbf{t}} = B \circ \mathbf{t}$. This implies that \bar{c} differs from $B \circ c$ by a translation. ♦

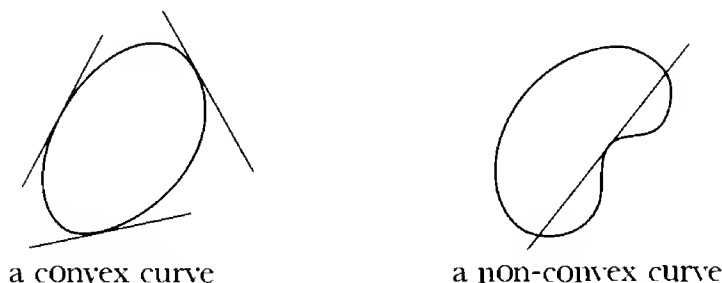
Notice that this theorem makes the previous calculation unnecessary: Since a circle of radius R has constant curvature $1/R$, any curve with constant curvature $1/R$ differs from this circle by a Euclidean motion, and is consequently another circle of radius $1/R$. More generally, Theorem 2 seems to make further study of curves almost pointless. Although one may still study properties of curvature, there is clearly no point in introducing any similar concept; we would only be interested in concepts that remained the same for c and $A \circ c$, and all of these are already determined by the curvature.

Despite these remarks, we are by no means ready to write off the study of plane curves. Many interesting results remain, of which we will be able to sample only a few. However, these results were all proved many years after the study of curves had been initiated, and are all global results, rather than local ones. To begin, we define certain kinds of curves, with which we will be almost exclusively concerned.

A C^1 curve $c: [a, b] \rightarrow \mathbb{R}^2$ is called **closed** if $c(a) = c(b)$ and $c'(a) = c'(b)$.



One can also regard a closed curve as an immersion of S^1 in \mathbb{R}^2 . A curve is called **simple** if it is one-one. Finally, among the simple closed curves we distinguish a special class of curves called **convex**. These are defined to be the simple closed curves which always lie on one side of their tangent lines.

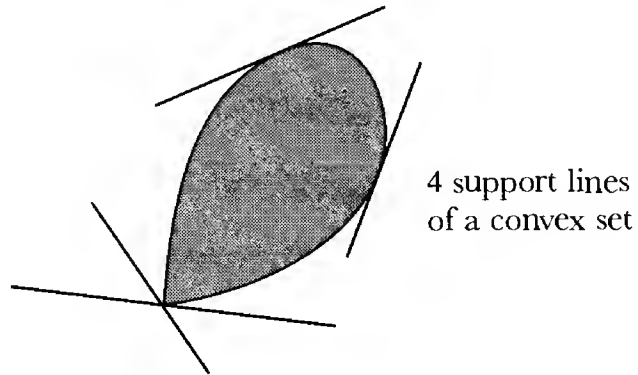


Although the property enunciated in this definition is precisely the one which is used in all proofs about convex curves, we will nevertheless take time out to equate this definition with a more common one.

Any subset A of \mathbb{R}^2 is called **convex** if the line segment \overline{pq} from p to q is

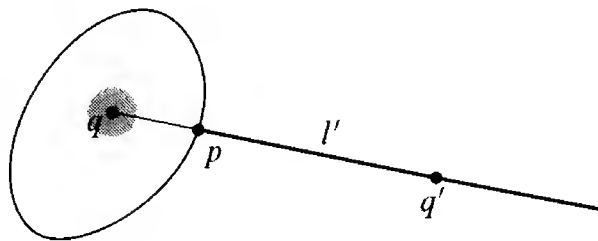


contained in A whenever $p, q \in A$. Suppose A is convex and p is a point in the boundary of A . A line L through p is called a **support line** of A if A lies completely in one of the closed half-spaces into which L divides \mathbb{R}^2 .



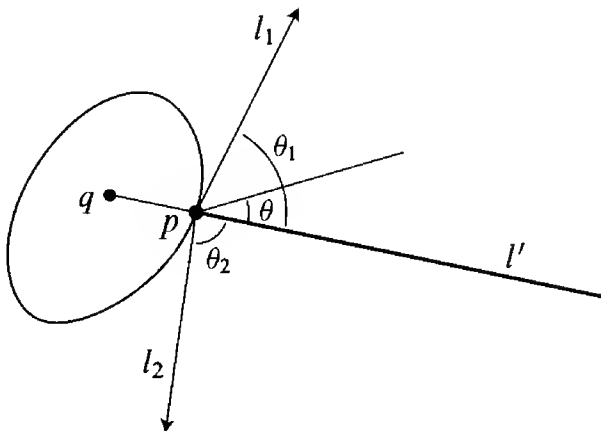
3. PROPOSITION. If A is convex, and p is in the boundary of A , then there is at least one support line L through p .

PROOF. If A has no interior points it lies on a line, and the proof is trivial. If A has an interior point q , let l be the ray from q through p , and let l' be the part starting at p . Clearly l' intersects A only at p , for if a point q' on l' were

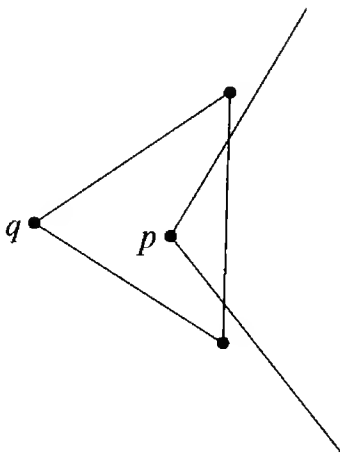


in A , then all points between q' and the points in a neighborhood of q would be in A , so p would be an interior point of A .

Choose one side of l' , and consider angles θ such that rays from p making an angle of θ with l' on this side do not intersect A except at p (it may be that $\theta = 0$ is the only possibility). Let θ_1 be the least upper bound of all such θ , and let l_1 be the ray through p making an angle of θ_1 . Let l_2 be the corresponding ray on the other side of l' .



We claim that the angle between l_1 and l_2 is $\geq \pi$, which will surely prove the theorem. To prove the claim, note that there are points of A arbitrarily close to (or perhaps even on) both l_1 and l_2 . If the angle between l_1 and l_2 were $< \pi$,



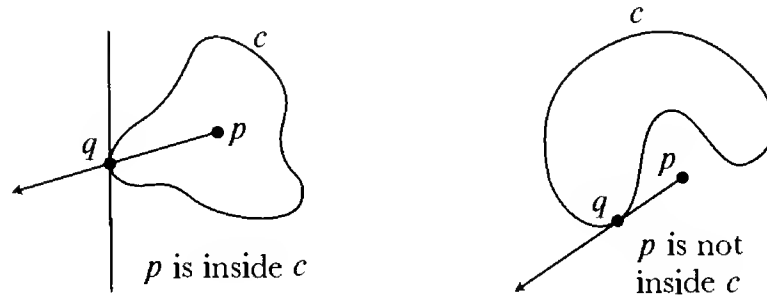
the triangle containing a suitable pair of such points, and q , would contain p in its interior, which cannot happen, since p is not an interior point of A . ❖

Note (for those who are familiar with Banach spaces). This theorem is essentially the Hahn-Banach Theorem. If A were symmetric about the origin, then \bar{A} would be the unit ball in \mathbb{R}^2 for a Banach space norm $\| \cdot \|$. If $W \subset \mathbb{R}^2$ is the subspace spanned by p , and $\lambda: W \rightarrow \mathbb{R}$ is the linear functional with $\lambda(p) = 1$, then the

desired support line is just a translate of the kernel of an extension $\bar{\lambda}: \mathbb{R}^2 \rightarrow \mathbb{R}$ of λ with $\|\bar{\lambda}\| = 1$. Symmetry of A is really unimportant, for the Hahn-Banach Theorem only requires a norm satisfying $\|av\| = a\|v\|$ for $a > 0$. The proof given here is just a geometrical translation of the main step in the usual proof of the Hahn-Banach Theorem.

We now want to show that a simple closed curve c is convex if and only if the set A consisting of all points on c or inside c is a convex subset of \mathbb{R}^2 . This is going to be pretty hard, since we have never defined the inside of a simple closed curve, and are just assuming that the content of Corollary I.11-15 is intuitively obvious. There is really no need to go through the proof of all this right now; it is only necessary to accept the following fact:

Suppose c is a simple closed curve, and l is a ray from p which intersects c at just one point $q \neq p$. Suppose, moreover, that the tangent line of c at q does *not* lie along l . Then p is inside c .

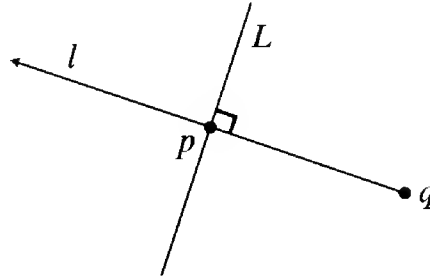


4. PROPOSITION. Let c be a simple closed curve, and let A be the set of all points on or inside c . Then c is convex (that is, c lies on one side of each of its tangent lines) if and only if A is convex.

PROOF. Suppose first that A is convex. Any point p on c is a boundary point of A , so there is a support line L of A through p . This line is clearly the tangent line of c at p , so c lies on one side of the tangent line through p .

Now suppose c is convex. For each p on c , let H_p be the closed half-plane, bounded by the tangent line through p , in which c lies. Clearly $A \subset \bigcap_p H_p$. Since the intersection $\bigcap_p H_p$ of all H_p is convex (any intersection of convex sets is convex), it suffices to show that we actually have $A = \bigcap_p H_p$. So consider a

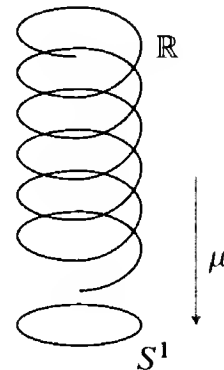
point q which is outside A . Let p be the point on c closest to q . Clearly the tangent line L of c at p is perpendicular to the ray l from q through p . It



suffices to show that c lies on the opposite side of L from q , for then $q \notin H_p$. Now if c lay entirely on the same side of L as q , then c could not intersect the ray l at any other point; for it cannot intersect the open segment \overline{pq} , since p is the closest point on c to q , and it certainly could not intersect l at points on the other side of L . By the remark preceding the Proposition, this would mean that q is inside c , a contradiction. ♦

Our first global results about curves depend upon a corresponding global formulation of the curvature function for a curve $c: [a, b] \rightarrow \mathbb{R}^2$. As usual, we assume c is parameterized by arclength, and consider its associated unit tangent vector curve $\mathbf{t}: [a, b] \rightarrow S^1$. Any point in S^1 can be described as $(\cos \theta, \sin \theta)$ for a real number θ . Of course, it is not possible to do this continuously. More precisely, if we define $\mu: \mathbb{R} \rightarrow S^1$ by

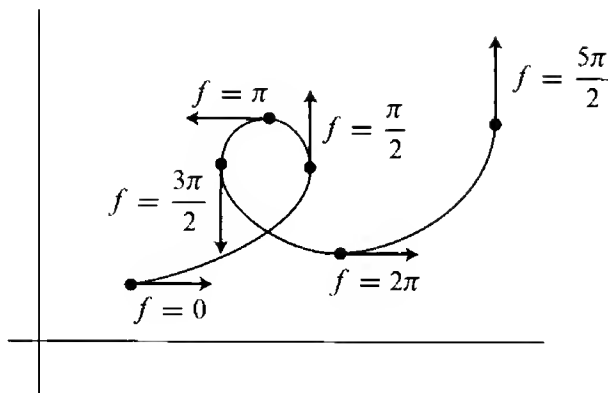
$$\mu(\theta) = (\cos \theta, \sin \theta) \in S^1,$$



then there is no continuous function $f: S^1 \rightarrow \mathbb{R}$ with $\mu \circ f = \text{identity}$. On the other hand, for our curve $c: [a, b] \rightarrow \mathbb{R}^2$ there is a continuous function

$f: [a, b] \rightarrow \mathbb{R}$ with

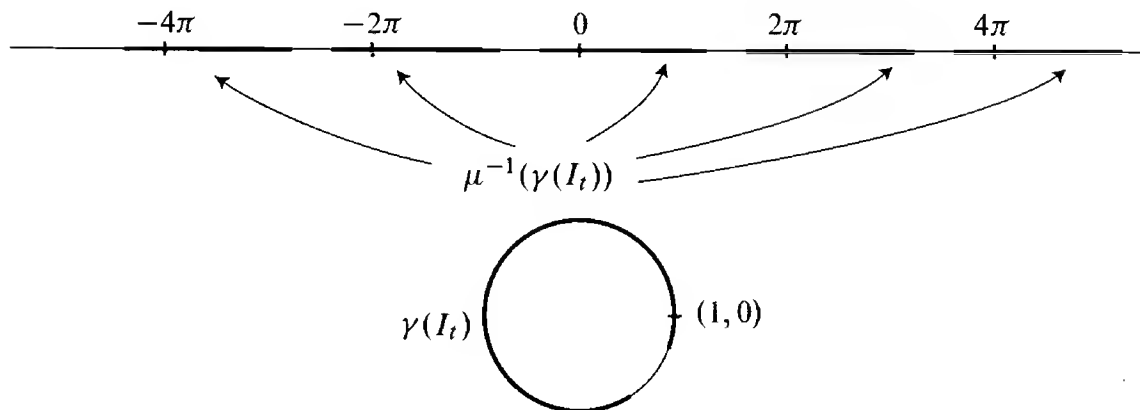
$$\mu(f(s)) = \mathbf{t}(s).$$



More generally,

5. PROPOSITION. Let $\gamma: [a, b] \rightarrow S^1$ be continuous. Then there is a continuous function $f: [a, b] \rightarrow \mathbb{R}$ with $\mu \circ f = \gamma$. Moreover, if f and \tilde{f} are any two such functions, then $f - \tilde{f} = 2\pi k$ for some k .

PROOF. For any $t \in [a, b]$ there is an open connected subset I_t of $[a, b]$ containing t such that $\gamma(I_t)$ is a *proper* connected subset of S^1 . Clearly $\mu^{-1}(\gamma(I_t))$ is then a disjoint union of connected sets and μ restricted to each of them is a homeomorphism onto $\gamma(I_t)$. This shows that if we choose any number v with



$\mu(v) = \gamma(t)$, then f can be defined uniquely on I_t in such a way that it has the value v at t , is continuous, and satisfies $\mu \circ f = \gamma$ on I_t .

We first prove that if there are two continuous functions f and \tilde{f} with $\mu \circ f = \gamma$, then they must differ by $2\pi k$ for some k . It obviously suffices to prove that $f = \tilde{f}$ if $f(a) = \tilde{f}(a)$. Let A be the set of all $t \in [a, b]$ such that $f(t) = \tilde{f}(t)$. Then A is closed, while the previous paragraph shows that A is open. Since $a \in A$ and $[a, b]$ is connected, this shows that $A = [a, b]$.

To prove existence, consider the set B of all $t \in [a, b]$ such that a continuous f can be defined on $[a, t]$, and let t_0 be the least upper bound of B . There is $t_1 \in B$ with $t_1 < t_0$ and $t_1 \in I_{t_0}$. Define a continuous \tilde{f} on I_{t_0} (with any value v at t_0 satisfying $\mu(v) = \gamma(t_0)$). Then $\tilde{f}(t_1) - f(t_1) = 2k\pi$ for some k . Now $\tilde{f} - 2k\pi$ must equal f on $[t_1, t_0]$, by the uniqueness proved previously. So we can extend f to be $\tilde{f} - 2k\pi$ on $[0, t_0]$. This shows that $t_0 \in B$. If $t_0 < b$, then we obtain an immediate contradiction by extending f to $[a, t_0] \cup I_{t_0}$. ♦

If f is the function given by Proposition 5 for the curve $\mathbf{t}: [a, b] \rightarrow S^1$, so that

$$c'(s) = \mathbf{t}(s) = \mu(f(s)) = (\cos f(s), \sin f(s)),$$

then

$$c''(s) = (-f'(s) \sin f(s), f'(s) \cos f(s)).$$

So

$$\begin{aligned} \kappa(s) &= c_1'(s)c_2''(s) - c_2'(s)c_1''(s) \\ &= \cos f(s) \cdot [(f'(s) \cos f(s))] - \sin f(s) \cdot [-f'(s) \sin f(s)]. \end{aligned}$$

We thus have

$$\kappa(s) = f'(s).$$

Notice that this gives us an easy way to reconstruct the curve from its curvature function: We first reconstruct f as $f(s) = \int_0^s \kappa(\theta) d\theta$; this gives us $\mathbf{t}(s) = \mu(f(s))$, so one more integration gives us the curve. Notice also that

$$\begin{aligned} f(b) - f(a) &= \int_a^b f'(s) ds \\ &= \int_a^b \kappa(s) ds; \end{aligned}$$

this quantity is called the **total curvature** of c . If c is a closed curve, the function $\mathbf{t}: [a, b] \rightarrow S^1$ satisfies $\mathbf{t}(a) = \mathbf{t}(b)$, so we may regard it as a map $\mathbf{t}: S^1 \rightarrow S^1$. The total curvature then has a special interpretation.

6. PROPOSITION. The total curvature of a closed curve $c: [a, b] \rightarrow \mathbb{R}^2$ is 2π times the degree of the map $\mathbf{t}: S^1 \rightarrow S^1$. (The degree of a map is defined on pg. I.275.)

PROOF. Since the form “ $d\theta$ ” on S^1 has integral 2π , the degree of \mathbf{t} is

$$\begin{aligned}\frac{1}{2\pi} \int_a^b \mathbf{t}^*(d\theta) &= \frac{1}{2\pi} \int_a^b (\mu \circ f)^*(d\theta) \\ &= \frac{1}{2\pi} \int_a^b f^*(\mu^*(d\theta)).\end{aligned}$$

Now

$$\mu_* \left(\frac{d}{dt} \Big|_t \right) = (-\sin \theta, \cos \theta)_{\mu(\theta)} \in S^1_{\mu(\theta)};$$

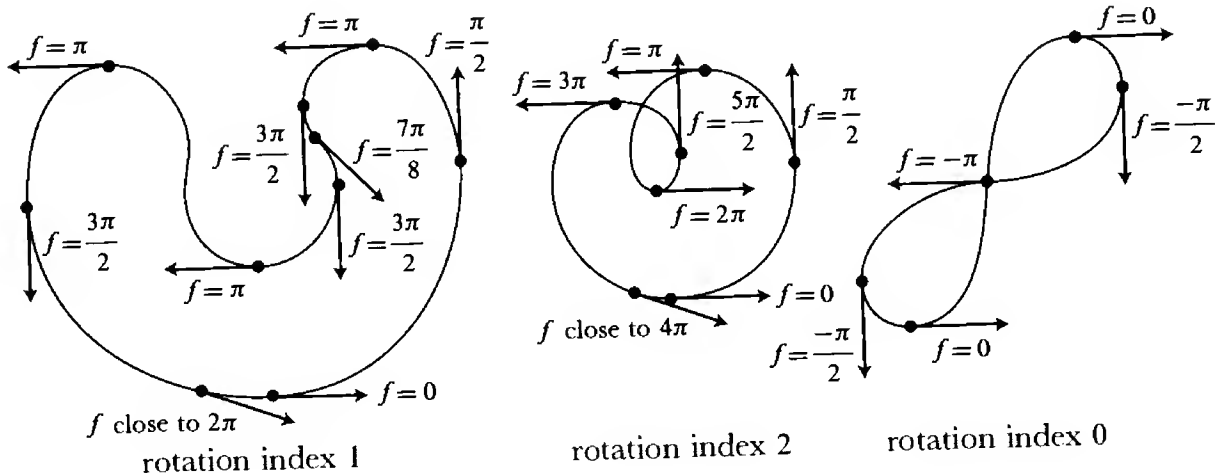
this is a unit tangent vector of S^1 , on which $d\theta$ has the value 1. So

$$\mu^*(d\theta) = dt.$$

Thus

$$\begin{aligned}\text{degree of } \mathbf{t} &= \frac{1}{2\pi} \int_a^b f^*(dt) \\ &= \frac{1}{2\pi} \int_a^b f' dt \\ &= \frac{1}{2\pi} [f(b) - f(a)]. \quad \spadesuit\end{aligned}$$

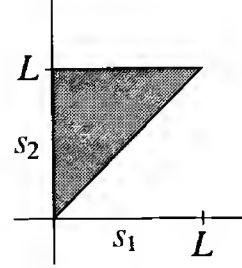
The degree of \mathbf{t} is also called the **rotation index** of c . The left-most figure below illustrates the first of our global theorems.



7. THEOREM (THE HOPF UMLAUFSATZ). The rotation index of a simple closed curve is ± 1 (depending on the direction in which it is traversed).

PROOF. Let $c: [0, L] \rightarrow \mathbb{R}^2$ be the curve, parameterized by arclength, and let $\Delta \subset \mathbb{R}^2$ be

$$\Delta = \{(s_1, s_2) : 0 \leq s_1 \leq s_2 \leq L\}.$$



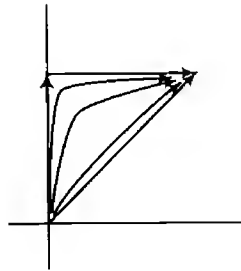
We define $\phi: \Delta \rightarrow S^1$ by

$$\phi(s_1, s_2) = \frac{c(s_2) - c(s_1)}{|c(s_2) - c(s_1)|} \quad s_1 < s_2 \quad \text{and} \quad (s_1, s_2) \neq (0, L)$$

$$\phi(s, s) = c'(s) = \mathbf{t}(s)$$

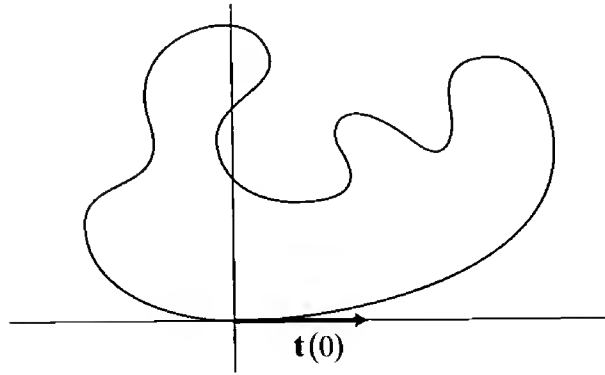
$$\phi(0, L) = -c'(0) = -\mathbf{t}(0).$$

It is easy to see that ϕ is continuous. Now the map $s \mapsto \phi(s, s)$ is just $\mathbf{t}: [0, L] \rightarrow S^1$. On the other hand, this map is homotopic to the map γ obtained by applying ϕ to the curve which goes along the other two sides of the triangle, from $(0, 0)$ to (L, L) . So it suffices to compute the degree of γ , which we break



into two pieces γ_1 and γ_2 , defined on $[0, L]$ and $[L, 2L]$, say.

The rotation index of c clearly does not change if we rotate or translate c , so we can assume that c lies in the upper half-plane, with $c(0) = (0, 0)$, and that the tangent line at $c(0)$ is the x -axis, as in the picture at the top of the next page. We assume that c is traversed in such a way that $\mathbf{t}(0) = (1, 0)$. Now, $\gamma_1(s) = \phi(0, s)$ clearly always lies in the semi-circle in the upper half-plane, and $\gamma_1(0) = \mathbf{t}(0)$, while $\gamma_1(L) = -\mathbf{t}(0)$. Consequently, the function f given by Proposition 5 clearly has its image in $[0, \pi]$, with $f(0) = 0$ and $f(L) = \pi$. Similarly, γ_2 lies in the lower half-plane, so the f for γ_2 satisfies $f(L) = \pi$ and $f(2L) = 2\pi$. Thus the degree of \mathbf{t} is $1/2\pi \cdot [2\pi - 0] = 1$. ♦

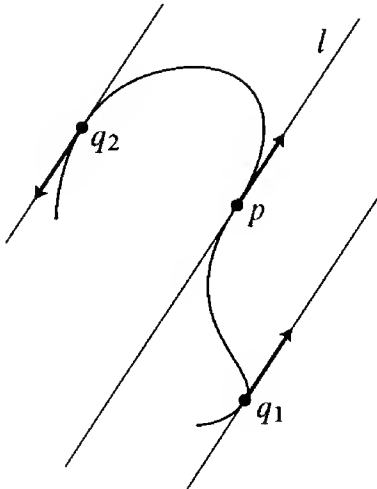


Using Theorem 7, we can now relate convexity and curvature.

8. THEOREM. A simple closed curve c is convex if and only if its curvature κ satisfies $\kappa \geq 0$ or $\kappa \leq 0$ (depending on the direction in which c is traversed).

PROOF. Let $c: [0, L] \rightarrow \mathbb{R}^2$ be parameterized by arclength, and choose a continuous $f: [0, L] \rightarrow \mathbb{R}$ with $\mu \circ f = \mathbf{t}$.

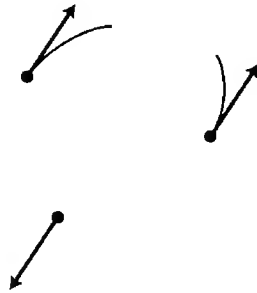
If $\kappa \geq 0$, then $f' \geq 0$, so f is non-decreasing. Suppose that c were not convex, so that c lies on both sides of the tangent line l through some point p . There are points q_1, q_2 on c which are furthest away from l on both sides of l .



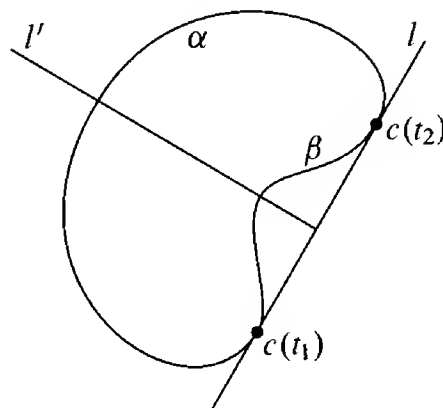
The tangent lines at q_1 and q_2 are clearly parallel to l , so of the three unit tangents at these points, at least two are identical, say $\mathbf{t}(s_1) = \mathbf{t}(s_2)$ for $s_1 < s_2$, with $(s_1, s_2) \neq (0, L)$. Thus $f(s_2) - f(s_1)$ is a multiple of 2π . But f is non-decreasing, and $f(L) = f(0) + 2\pi$, by Theorem 7. So either $f(s_2) = f(s_1)$ or $f(s_2) = f(s_1) + 2\pi$. In the first case it follows that f is constant on $[s_1, s_2]$, so \mathbf{t} is

constant on $[s_1, s_2]$. This implies that c is a straight line on $[s_1, s_2]$, contradicting the fact that none of the points p, q_1, q_2 lies on the tangent lines through the others. If $f(s_2) = f(s_1) + 2\pi$, the other arc of c must similarly be a straight line, which is again a contradiction. Thus c must be convex.

Now suppose c is convex. If there is $s_1 < s_2$ with $f(s_1) = f(s_2)$, so that $\mathbf{t}(s_1) = \mathbf{t}(s_2)$, then there is also s with $\mathbf{t}(s) = -\mathbf{t}(s_1)$, since \mathbf{t} has degree 1, and



is consequently onto S^1 . The tangent lines through two of the three points $c(s), c(s_1), c(s_2)$ must coincide (otherwise c would cross one of them). Thus c is tangent to the same line l at two points, $c(t_1)$ and $c(t_2)$, say. These two points divide c into two arcs, α and β . If l' is perpendicular to l , and intersects l



between $c(t_1)$ and $c(t_2)$, then α and β must intersect l' ; using convexity of c , it is easy to see that α and β intersect l' exactly once. Clearly one arc, say α , always intersects l' at a point further away from l than the other, β . We claim that β lies along l . If not, consider the tangent line through the point P of β furthest from l . This tangent line is parallel to l , and c would lie on both sides of it, a contradiction.

Now, since β lies along l , we have $\mathbf{t}(t_1) = \mathbf{t}(t_2)$. Thus $t_1 = s_1$ and $t_2 = s_2$. Since the curve lies along a line on $[s_1, s_2]$, we have $f(s) = f(s_1)$ for all $s \in [s_1, s_2]$. Thus, f is non-decreasing. ♦

Our final global result about curves in the plane involves another local concept. A **vertex** of a curve c is a point where $\kappa'(s) = 0$. It is easily seen that an ellipse which is not a circle has exactly four vertices, at the ends of the major and minor axes; these are the points where κ has a local maximum or minimum (though a vertex need not generally be of this type). Before proceeding with the next theorem we need a preliminary observation. The curvature κ is defined by

$$\mathbf{t}'(s) = \kappa(s) \cdot \mathbf{n}(s);$$

the existence of such a number $\kappa(s)$ follows from the equation $\langle \mathbf{t}, \mathbf{t} \rangle = 1$, by differentiation. Similarly, from $\langle \mathbf{n}, \mathbf{n} \rangle = 1$ we obtain

$$\langle \mathbf{n}'(s), \mathbf{n} \rangle = 0,$$

which implies that $\mathbf{n}'(s)$ is a multiple of $\mathbf{t}(s)$, say $\mathbf{n}'(s) = \alpha \cdot \mathbf{t}(s)$. On the other hand, from $\langle \mathbf{t}, \mathbf{n} \rangle = 0$ we obtain

$$\langle \mathbf{t}'(s), \mathbf{n}(s) \rangle + \langle \mathbf{t}(s), \mathbf{n}'(s) \rangle = 0,$$

or

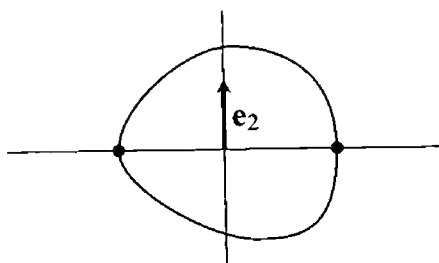
$$\kappa(s) + \alpha = 0;$$

hence,

$$(*) \quad \mathbf{n}'(s) = -\kappa(s) \cdot \mathbf{t}.$$

9. THEOREM (THE FOUR VERTEX THEOREM). Every simple closed convex curve has at least four vertices.

PROOF. If $c: [0, L] \rightarrow \mathbb{R}^2$ is the simple closed curve, parameterized by arc-length, then c has at least two vertices—namely, the maximum and minimum points for the curvature. Choose the coordinate system so that the x -axis passes through these two points. Now integration by parts gives the following equation,



in which we are taking integrals of \mathbb{R}^2 -valued functions by integrating each component separately:

$$\begin{aligned}\int_0^L \kappa'(s) \cdot c(s) &= - \int_0^L \kappa(s) \cdot \mathbf{t}(s) ds \\ &= \int_0^L \mathbf{n}'(s) ds \quad \text{by } (*) \\ &= \mathbf{n}(L) - \mathbf{n}(0) = 0.\end{aligned}$$

Consequently, we certainly have

$$(**) \quad \int_0^L \kappa'(s) \langle c(s), e_2 \rangle ds = 0.$$

If there are no other vertices, then $\kappa' > 0$ on one half of c and $\kappa' < 0$ on the other. So $\kappa'(s) \langle c(s), e_2 \rangle$ has the same sign on both halves, contradicting (**). Thus c must have at least one more vertex.

The argument just given actually shows that c cannot be formed of two arcs with $\kappa' > 0$ on one and $\kappa' < 0$ on the other; the same conclusion clearly holds even if $\kappa' \geq 0$ on one and $\kappa' \leq 0$ on the other, since $\kappa' \neq 0$ somewhere. This shows that c must have a fourth vertex; if it had only three, then some pair would divide c into two arcs with $\kappa \geq 0$ on one and $\kappa' \leq 0$ on the other. ♦

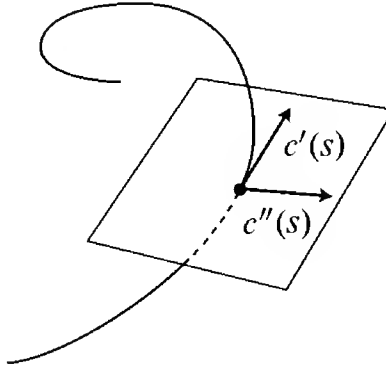
We now turn our attention to curves in space, and ask, once again, how to measure the curvature of $c: [a, b] \rightarrow \mathbb{R}^3$. We can still look for the limiting position of circles through $c(s_1), c(s_2), c(s_3)$ as $s_1, s_2, s_3 \rightarrow s$. If this limiting circle exists, its center C must satisfy

$$\begin{aligned}\langle c'(s), c(s) - C \rangle &= 0 \\ \langle c''(s), c(s) - C \rangle &= -\langle c'(s), c'(s) \rangle;\end{aligned}$$

the derivation of this necessary condition still works. However, for space curves these equations do not even determine C ; the first equation merely restricts C to lie on a certain plane, not on a certain line. We must first see whether the *planes* through the points $c(s_i)$ approach a limiting position.

10. PROPOSITION. Let $c: [a, b] \rightarrow \mathbb{R}^3$ be a C^2 curve parameterized by arc-length, with $c''(s) \neq 0$. For s_1, s_2, s_3 sufficiently close to s , the points $c(s_1), c(s_2), c(s_3)$ do not lie on a line. As $s_i \rightarrow s$, the unique plane through the points $c(s_i)$ approaches the plane P spanned by $c'(s)$ and $c''(s)$.

Remark: The plane P should really be described as the plane through $c(s)$ which is parallel to the plane spanned by $c'(s)$ and $c''(s)$, but we will allow ourselves the elliptical terminology suggested by the picture.



PROOF. Assuming for the moment that the points $c(s_i)$ do not lie on a straight line, let $P(s_1, s_2, s_3)$ be the plane spanned by these points, and let $a(s_1, s_2, s_3)$ be a unit vector perpendicular to $P(s_1, s_2, s_3)$. Then the function

$$(*) \quad s \mapsto \langle c(s), a(s_1, s_2, s_3) \rangle$$

is 0 for $s = s_i$. So we have

$$(1) \quad \langle c'(\xi_i), a(s_1, s_2, s_3) \rangle = 0 \quad \xi_i \in (s_i, s_{i+1}), \quad i = 1, 2.$$

It follows that

$$(2) \quad \langle c''(\eta), a(s_1, s_2, s_3) \rangle = 0 \quad \eta \in (\xi_1, \xi_2).$$

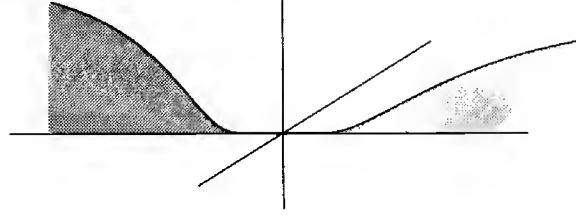
Equations (1) and (2), together with continuity of c' and c'' , clearly show that $a(s_1, s_2, s_3)$ approaches a unit vector perpendicular to $c'(s)$ and $c''(s)$, so that $P(s_1, s_2, s_3)$ approaches P .

If the points $c(s_i)$ do lie on a straight line, then we can choose a whole circle of unit vectors $a(s_1, s_2, s_3)$ for which the function $(*)$ vanishes at $s = s_i$. If this were true for s_i arbitrarily close to s , the remaining part of the argument would imply that all of these vectors are nearly perpendicular to P , which is absurd. ♦

The plane described in Proposition 10 is called the **osculating plane** of the curve at s . Notice that, unlike the osculating circle, the osculating plane may exist even if $c''(s) = 0$. For example, if c is a plane curve which is not straight, then the osculating plane certainly exists. The exact conditions for the existence of an osculating plane are not very important for us, but we will pause to indicate the actual state of affairs.

The worst possible situation occurs for the curve

$$c(t) = \begin{cases} 0 & t = 0 \\ (t, e^{-1/t^2}, 0) & t > 0 \\ (t, 0, e^{-1/t^2}) & t < 0, \end{cases}$$



which can't make up its mind whether to osculate in the (x, y) -plane or in the (x, z) -plane. For this curve we have $c^{(k)}(0) = 0$ for all $k \geq 2$, which suggests that we consider only curves with $c^{(k)}(s) \neq 0$ for some $k \geq 2$. Notice that for curves parameterized by arclength, $\langle c'(s), c''(s) \rangle = 0$ implies

$$\langle c'(s), c'''(s) \rangle + \langle c''(s), c''(s) \rangle = 0.$$

So if $c''(s) = 0$, we have $c'''(s)$ perpendicular to $c'(s)$. Similarly, $c'''(s) = 0$ implies that $c^{(4)}(s)$ is perpendicular to $c'(s)$, etc. So the first non-zero higher derivative $c^{(k)}(s)$ is the same as the first derivative which is linearly independent of $c'(s)$. Now suppose P is the plane spanned by $c'(s)$ and this first non-zero $c^{(k)}(s)$. Let s_1, s_2, s_3 be parameter values *on the same side* of s , i.e., $s \leq s_1 < s_2 < s_3$ (or $s_1 < s_2 < s_3 \leq s$). If $a(s_1, s_2, s_3)$ is a unit vector perpendicular to all $c(s_i) - c(s)$, then, as before, we have

$$\begin{aligned} \langle c'(\xi_i), a(s_1, s_2, s_3) \rangle &= 0 \\ \langle c''(\eta), a(s_1, s_2, s_3) \rangle &= 0. \end{aligned}$$

Now $\eta \in (s_1, s_3)$ so $\eta \neq s$. Hence, if $c''(s) = 0$ we have $\theta \in (s, \eta)$ with

$$\langle c'''(\theta), a(s_1, s_2, s_3) \rangle = 0.$$

Continuing in this way, we finally obtain $\lambda \in (s, s_3)$ with

$$\langle c^{(k)}(\lambda), a(s_1, s_2, s_3) \rangle = 0.$$

As before, this shows that the plane through the points $c(s_i)$ approaches P , and that the points $c(s_i)$ cannot lie on a line for s_i arbitrarily close to 0.

If the s_i are allowed to lie on both sides of s , this result is no longer true. For example, consider the curve

$$c(t) = (t, t^4, t^3).$$

We have

$$\begin{aligned}c'(0) &= (1, 0, 0) \\c''(0) &= (0, 0, 0) \\c'''(0) &= (0, 0, 6),\end{aligned}$$

so $(0, 1, 0)$ is perpendicular to the plane spanned by $c'(0)$ and $c'''(0)$. On the other hand, for a vector perpendicular to the plane spanned by $c(0), c(t), c(-t)$ we can choose

$$\begin{aligned}\text{normalized } [c(t) - c(0)] \times [c(-t) - c(0)] \\&= \text{normalized } (t, t^4, t^3) \times (-t, t^4, -t^3) \\&= \text{normalized } (-2t^7, 0, 2t^5) \\&\rightarrow (0, 0, 1).\end{aligned}$$

This strange behavior is clarified by a look at the Taylor expansion

$$c(s) = c(0) + sc'(0) + 0 + \frac{s^3}{6}c'''(0) + o(s^3),$$

which shows that

$$\begin{aligned}[c(s) - c(0)] \times [c(t) - c(0)] &= \left[\frac{st^3}{6} - \frac{ts^3}{6} \right] c'(0) \times c'''(0) \\&\quad + \text{higher order terms.}\end{aligned}$$

When s and t have the same sign, the dominant term is the first. But when s and t have opposite signs this is no longer true—this term may even be 0. I suspect, but have not checked, that the parameter values s_i may be picked on both sides of 0 if and only if the first $k \geq 2$ with $c^{(k)}(0) \neq 0$ is even.

For space curves $c: [a, b] \rightarrow \mathbb{R}^3$ with $c''(s) \neq 0$, we now clearly have an osculating circle, the limit as $s_1, s_2, s_3 \rightarrow s$ of the circle through the points $c(s_i)$; it lies in the osculating plane. We define the curvature κ to be the reciprocal of the radius of this circle, so that

$$\kappa(s) = |\mathbf{t}'(s)|.$$

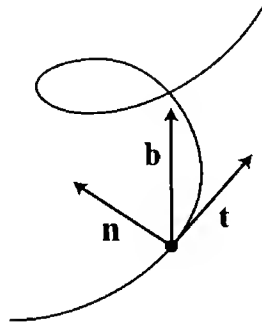
This definition, once again, determines a curvature even when $c''(s) = \mathbf{t}'(s) = 0$. Unlike the case of plane curves, we cannot obtain a signed curvature, for there is no natural way to pick a vector orthogonal to $\mathbf{t}(s)$. However, when $\kappa(s) \neq 0$, we can now *define* $\mathbf{n}(s)$ by the equation

$$\mathbf{t}'(s) = \kappa(s)\mathbf{n}(s), \quad \mathbf{n}(s) = \text{normalized } \mathbf{t}'(s).$$

The vector $\mathbf{n}(s)$ is called the **principal normal** of c at s . We also define the **binormal** $\mathbf{b}(s)$ by

$$\mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s),$$

so that $(\mathbf{t}, \mathbf{n}, \mathbf{b})$ is always a positively oriented orthonormal basis.



Note that $\langle \mathbf{b}, \mathbf{b} \rangle = 1$ implies that $\langle \mathbf{b}', \mathbf{b} \rangle = 0$, so that \mathbf{b}' is a linear combination of \mathbf{t} and \mathbf{n} . We also have $\langle \mathbf{b}, \mathbf{t} \rangle = 0$, which implies that

$$\langle \mathbf{b}', \mathbf{t} \rangle = -\langle \mathbf{b}, \mathbf{t}' \rangle = -\langle \mathbf{b}, \mathbf{n} \rangle = 0.$$

Thus \mathbf{b}' is actually a multiple of \mathbf{n} , and we can define a new function τ , the **torsion**, by

$$\mathbf{b}' = -\tau \mathbf{n}.$$

Of course, we can define τ only at points where \mathbf{n} exists, i.e., where $c'' \neq 0$. This is analogous to the fact that κ can be defined only at points where $c' \neq 0$. The reason for choosing the negative sign in this equation will be explained in a moment; we first interpret the absolute value $|\tau|$. The function $\mathbf{b}: [a, b] \rightarrow S^2$ has an arclength function

$$\begin{aligned} \text{length of } \mathbf{b} \text{ on } [a, s] &= \int_a^s |\mathbf{b}'(u)| du \\ &= \int_a^s |\tau(u)| du. \end{aligned}$$

Consequently, $|\tau(s)|$ is the derivative of this arclength function. Since \mathbf{b} is the perpendicular to the osculating plane, this derivative of the length of \mathbf{b} can be thought of as *the rate at which the osculating plane is changing*. Thus $|\tau|$ measures, in some sense, the rate at which the curve deviates from being a plane curve. Classically, curvature and torsion were also known as *first* and *second curvature*, and space curves were called *curves of double curvature*.

To develop a formula for the torsion, we first recall that the cross product $v \times w$ of v and w is defined by the equation

$$(*) \quad \langle z, v \times w \rangle = \det \begin{pmatrix} z \\ v \\ w \end{pmatrix}.$$

[Since the i^{th} component of $v \times w$ is $\langle e_i, v \times w \rangle$, this shows that $v \times w$ can be obtained by computing the determinant

$$v \times w = \det \begin{pmatrix} e_1 & e_2 & e_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} = (v_2 w_3 - w_2 v_3) e_1 + \cdots$$

purely formally.] Using the fact that \det is alternating, the left side of $(*)$, classically known as the “scalar triple product” (v, w, z) , is seen to satisfy

$$(**) \quad \langle z, v \times w \rangle = -\langle w, v \times z \rangle = \langle w, z \times v \rangle.$$

Now for the torsion τ of a curve c parameterized by arclength we clearly have

$$\begin{aligned} \tau &= \langle -\mathbf{n}, \mathbf{b}' \rangle = \langle -\mathbf{n}, (\mathbf{t} \times \mathbf{n})' \rangle \\ &= \langle -\mathbf{n}, \mathbf{t} \times \mathbf{n}' \rangle + \langle -\mathbf{n}, \mathbf{t}' \times \mathbf{n} \rangle, \quad \text{where } \mathbf{t}' \times \mathbf{n} = 0 \\ &= \left\langle -\frac{c''}{\kappa}, c' \times \left(\frac{c''}{\kappa} \right)' \right\rangle \\ &= \left\langle -\frac{c''}{\kappa}, c' \times \left(\frac{\kappa c''' - \kappa' c''}{\kappa^2} \right) \right\rangle \\ &= -\frac{1}{\kappa^2} \langle c'', c' \times c''' \rangle, \end{aligned}$$

or

$$\tau = \frac{1}{\kappa^2} \langle c' \times c'', c''' \rangle.$$

This shows that $\tau > 0$ when $\mathbf{t}, \mathbf{n}, c'''$ form a positively oriented basis for \mathbb{R}^3 , which is the same as saying that c''' is on the same side of the osculating plane as \mathbf{b} . Now in Taylor's formula,

$$c(s+h) = c(s) + hc'(s) + \frac{h^2}{2}c''(s) + \frac{h^3}{6}c'''(s) + o(h^3),$$

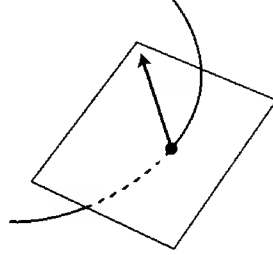
the term $c(s) + hc'(s) + h^2c''(s)/2$ is in the osculating plane at s , so

$$c(s+h) - \left[c(s) + hc'(s) + \frac{h^2}{2}c''(s) \right] = \frac{h^3}{6}c'''(s) + o(h^3)$$

points from the osculating plane to $c(s+h)$. Consequently, if $c'''(s) \neq 0$, then the curve pierces the osculating plane at s , and the points $c(s+h)$ for small

$h > 0$ are on the same side as $c'''(s)$, while points $c(s + h)$ for $h < 0$ are on the other side. Together with our previous remarks, this shows that

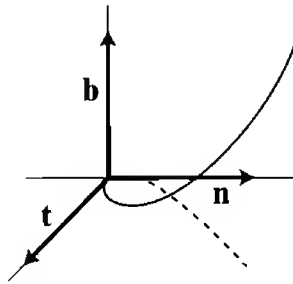
if $\tau(s) > 0$, then points $c(s + h)$ for small $h > 0$ lie on the same side of the osculating plane as $\mathbf{b}(s)$, while points $c(s + h)$ for small $h < 0$ lie on the opposite side.



Our formula for τ shows that the torsion for the curve $\bar{c} = s \mapsto c(-s)$ has the same sign as that of c . This is because reversing directions also reverses the binormal. For a curve with an arbitrary parameterization we obtain, proceeding just as in the case of curvature, the formula

$$\tau = \frac{\langle \dot{c} \times \ddot{c}, \ddot{\ddot{c}} \rangle}{\langle \dot{c} \times \ddot{c}, \dot{c} \times \ddot{c} \rangle}.$$

A standard and possibly illuminating way of examining the geometrical significance of κ and τ is to examine the projections of c on the planes spanned by any two of \mathbf{t} , \mathbf{n} , \mathbf{b} . The plane spanned by \mathbf{t} and \mathbf{n} is just the osculating plane.



The plane spanned by the principal normal \mathbf{n} and binormal \mathbf{b} is called, naturally enough, the **normal plane**, and the plane spanned by \mathbf{t} and \mathbf{b} is called the **rectifying plane** (this terminology is explained in Volume III, pg. 186). We can choose a coordinate system for \mathbb{R}^3 so that $c(0) = 0$ and so that the osculating plane of c at 0 is the (x, y) -plane. Further choosing $c'(0) = (1, 0, 0)$ we obtain

$$\begin{aligned} c(0) &= (0, 0, 0) \\ c'(0) &= (1, 0, 0) \\ c''(0) &= (0, \kappa, 0) \\ c'''(0) &= (_, _, \kappa\tau), \end{aligned}$$

the last equation following from the fact that $\tau\kappa^2 = \langle c' \times c'', c''' \rangle$, as shown above. The three components of the Taylor expansion

$$c(s) = c(0) + sc'(0) + \frac{s^2}{2}c''(0) + \frac{s^3}{6}c'''(0) + \cdots$$

give

$$c_1(s) = s + \text{terms of order 3 or more}$$

$$c_2(s) = \frac{\kappa}{2}s^2 + \text{terms of order 3 or more}$$

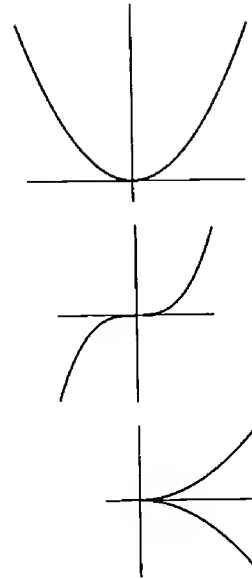
$$c_3(s) = \frac{\kappa\tau}{6}s^3 + \text{terms of order 4 or more.}$$

So the projections look like

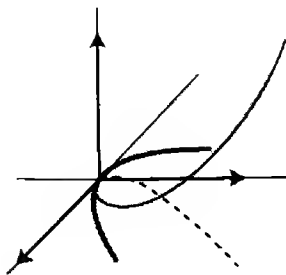
$$y = \frac{\kappa}{2}x^2 \text{ up to order 2 on the osculating plane}$$

$$z = \frac{\kappa\tau}{6}x^3 \text{ up to order 3 on the rectifying plane}$$

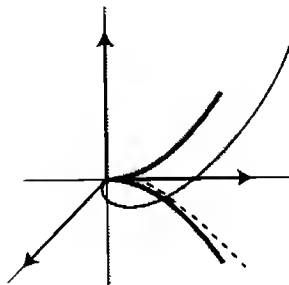
$$z^2 = \frac{2\tau^2}{9\kappa}y^3 \text{ up to order 3 on the normal plane.}$$



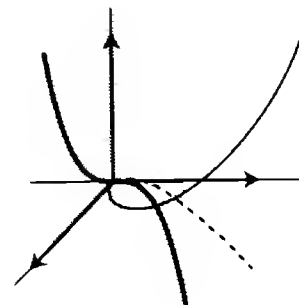
The figure below shows these projections for the curve on page 30.



osculating plane



normal plane



rectifying plane

Just as in the case of plane curves, we now ask to what extent the curvature and torsion determine a curve.

We note first that a curve c (parameterized by arclength) with $\kappa = 0$ everywhere is a straight line—the proof is the same as before.

Moreover, a curve with $\tau = 0$ everywhere is a plane curve. To prove this, we note that $\tau = 0$ means $\mathbf{b}' = 0$ so that $\mathbf{b}(s) = \mathbf{b}_0$, a constant vector. This implies that $\langle \mathbf{t}, \mathbf{b}_0 \rangle = 0$. But this means that

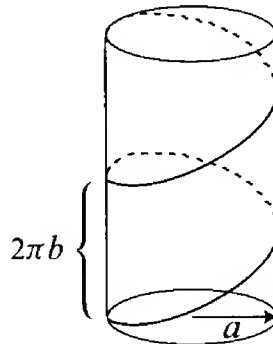
$$\frac{d}{ds} \langle c(s), \mathbf{b}_0 \rangle = 0,$$

so $\langle c(s), \mathbf{b}_0 \rangle = a$ where a is a constant. Thus c lies in the plane

$$\{p \in \mathbb{R}^3 : \langle p, \mathbf{b}_0 \rangle = a\}.$$

Unlike the case of plane curves, we should not expect a curve with constant curvature to be a circle, unless the torsion is 0. To get some idea of the variety of possibilities, we will examine only one special class of curves, the **helices**, given by the formula

$$c(u) = (a \cos u, a \sin u, bu).$$



We have

$$c'(u) = (-a \sin u, a \cos u, bu),$$

so

$$|c'(u)| = \sqrt{a^2 + b^2} = D,$$

and the reparameterization γ by arclength is given by

$$\gamma(s) = c(s/D),$$

with

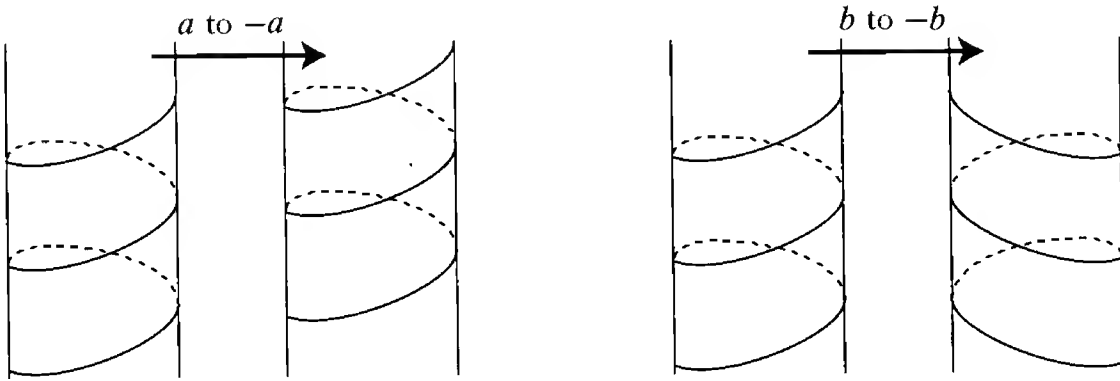
$$\begin{aligned} \gamma'(s) &= \left(-\frac{a}{D} \sin \frac{s}{D}, \frac{a}{D} \cos \frac{s}{D}, \frac{b}{D} \right) \\ \gamma''(s) &= \left(-\frac{a}{D^2} \cos \frac{s}{D}, -\frac{a}{D^2} \sin \frac{s}{D}, 0 \right). \end{aligned}$$

Thus

$$\kappa(s) = |\gamma''(s)| = \frac{|a|}{D^2},$$

$$\tau = \frac{\langle \gamma' \times \gamma'', \gamma''' \rangle}{\kappa^2} = \frac{b}{D^2}.$$

Notice that neither κ nor τ depend on the sign of a . Changing a to $-a$ merely rotates the helix through an angle of π around its axis, which is the same as moving it a certain distance in the direction of this axis. However, changing from b to $-b$ changes the helix from “right handed” to “left handed” or vice-versa, and accordingly changes the sign of τ .



By choosing suitable a and b , we can make $|a|/D^2$ and b/D^2 equal to any desired pair (κ, τ) with $\kappa > 0$. So helices give examples of curves with any desired constant curvature (> 0) and constant torsion. Are they the only such curves? Rather than imitating the calculations for the simpler question answered previously, we will immediately ask the more general question, whether κ and τ determine c up to a proper Euclidean motion.

We begin with a recapitulation of the definitions:

$$\mathbf{t}' = \kappa \mathbf{n}$$

$$\mathbf{b}' = -\tau \mathbf{n}.$$

Notice that we have expressed the derivatives of \mathbf{t} and \mathbf{b} in terms of the original vectors $\mathbf{t}, \mathbf{n}, \mathbf{b}$. We can do the same for \mathbf{n} . First, since $\langle \mathbf{n}, \mathbf{n} \rangle = 1$, we obtain $\langle \mathbf{n}', \mathbf{n} \rangle = 0$, so \mathbf{n}' is some linear combination of \mathbf{t} and \mathbf{b} . Now, from $\langle \mathbf{n}, \mathbf{t} \rangle = 0$ we obtain

$$\langle \mathbf{n}', \mathbf{t} \rangle = -\langle \mathbf{n}, \mathbf{t}' \rangle = -\langle \mathbf{n}, \kappa \mathbf{n} \rangle = -\kappa,$$

(we already obtained this equation for the case of plane curves); and from $\langle \mathbf{n}, \mathbf{b} \rangle = 0$ we obtain

$$\langle \mathbf{n}', \mathbf{b} \rangle = -\langle \mathbf{n}, \mathbf{b}' \rangle = -\langle \mathbf{n}, -\tau \mathbf{n} \rangle = \tau.$$

Thus we have, altogether,

$$\begin{array}{rcl} \mathbf{t}' & = & \kappa \mathbf{n} \\ \mathbf{n}' & = & -\kappa \mathbf{t} + \tau \mathbf{b} \\ \mathbf{b}' & = & -\tau \mathbf{n} \end{array}$$

These are called the **Serret-Frenet formulas**. They were obtained independently by Serret in 1851, and by Frenet, in his thesis of 1847, an abstract of which appeared in 1852. Before Serret and Frenet, many geometric properties of curves were investigated with great laboriousness, but afterwards many of these investigations became routine, because, as our next theorem shows, everything about space curves is contained in these formulas.

11. THEOREM. Let $\kappa, \tau: [a, b] \rightarrow \mathbb{R}$ be continuous, with $\kappa > 0$ on $[a, b]$. Then there is a curve $c: [a, b] \rightarrow \mathbb{R}^3$, parameterized by arclength, whose curvature and torsion functions are κ and τ . Any two such curves differ by a proper Euclidean motion (a translation followed by a rotation [an element of $\text{SO}(3)$]).

PROOF. Let us adopt the more systematic notation $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ for $\mathbf{t}, \mathbf{n}, \mathbf{b}$, and define a matrix

$$a_{ij}(s) = \begin{pmatrix} 0 & -\kappa(s) & 0 \\ \kappa(s) & 0 & -\tau(s) \\ 0 & \tau(s) & 0 \end{pmatrix},$$

so that the Serret-Frenet equations become

$$(*) \quad \mathbf{v}_i' = \sum_{j=1}^3 a_{ji} \mathbf{v}_j.$$

Now Theorem I.5-17 implies that there is a function $s \mapsto (\mathbf{v}_1(s), \mathbf{v}_2(s), \mathbf{v}_3(s))$ on $[a, b]$ satisfying (*). We can choose $\mathbf{v}_i(a)$ arbitrarily; choose them to be orthonormal and positively oriented. We claim that $\mathbf{v}_i(s)$ are orthonormal for all $s \in [a, b]$. This is the only significant point in the proof; the rest of the proof is exactly like the proof of Theorem 2.

To prove that the \mathbf{v}_i are always orthonormal, we note that

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle' = \sum_{k=1}^3 a_{ki} \langle \mathbf{v}_j, \mathbf{v}_k \rangle + a_{kj} \langle \mathbf{v}_i, \mathbf{v}_k \rangle.$$

This shows that the functions $\beta_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ satisfy the differential equation

$$(**) \quad \beta_{ij}' = \sum_{k=1}^3 a_{ki} \beta_{jk} + a_{kj} \beta_{ik},$$

together with the initial conditions $\beta_{ij}(a) = \delta_{ij}$. Since the solutions of (**) are determined by their initial conditions, we can prove \mathbf{v}_i everywhere orthonormal by showing that the functions $\beta_{ij}(s) = \delta_{ij}$ do satisfy (**). In other words, we want to show that

$$\begin{aligned} 0 = \delta_{ij}' &= \sum_{k=1}^3 a_{ki} \delta_{jk} + a_{kj} \delta_{ik} \\ &= a_{ji} + a_{ij}. \end{aligned}$$

But this is true—the matrix (a_{ij}) is skew-symmetric. ♦

The skew-symmetry of (a_{ij}) gives an easy way to remember the Serret-Frenet formulas: the formulas for \mathbf{t}' and \mathbf{b}' are by definition, and the formula for \mathbf{n}' is the one that makes the matrix skew-symmetric. If we think of $\mathbf{t}, \mathbf{n}, \mathbf{b}$ as column vectors, so that $(\mathbf{t}, \mathbf{n}, \mathbf{b})$ denotes a 3×3 matrix, then the Serret-Frenet equations read

$$(\mathbf{t}, \mathbf{n}, \mathbf{b})' = \begin{pmatrix} \mathbf{t}_1' & \mathbf{n}_1' & \mathbf{b}_1' \\ \mathbf{t}_2' & \mathbf{n}_2' & \mathbf{b}_2' \\ \mathbf{t}_3' & \mathbf{n}_3' & \mathbf{b}_3' \end{pmatrix} = \begin{pmatrix} \mathbf{t}_1 & \mathbf{n}_1 & \mathbf{b}_1 \\ \mathbf{t}_2 & \mathbf{n}_2 & \mathbf{b}_2 \\ \mathbf{t}_3 & \mathbf{n}_3 & \mathbf{b}_3 \end{pmatrix} \cdot \begin{pmatrix} 0 & -\kappa & 0 \\ \kappa & 0 & -\tau \\ 0 & \tau & 0 \end{pmatrix}.$$

Since $\mathbf{t}, \mathbf{n}, \mathbf{b}$ are orthonormal and positively oriented, the curve $\alpha(s) = (\mathbf{t}(s), \mathbf{n}(s), \mathbf{b}(s))$ is a curve in $\text{SO}(3)$, the identity component of the orthogonal group $\text{O}(3)$. Recall that the tangent space $\text{SO}(n)_I = \text{O}(n)_I = \mathfrak{o}(n)$ of $\text{SO}(n)$ at I is the set of skew-symmetric matrices. For $A \in \text{SO}(n)$, the tangent space $\text{SO}(n)_A$ is just $L_{A*}(\mathfrak{o}(n))$, which equals $L_A \cdot \mathfrak{o}(n)$, since L_A is a linear function. Thus $\text{SO}(n)_A$ consists of all matrices $A \cdot M$ for $M \in \mathfrak{o}(n)$. Clearly, $A \cdot M = L_{A*}(M)$ is just $\tilde{M}(A)$, where \tilde{M} denotes the left invariant vector field with value M at I . Now the curve $\alpha: [a, b] \rightarrow \text{SO}(3)$ must have its tangent vector $\alpha'(s)$ in $\text{SO}(3)_{\alpha(s)}$, so we must have

$$\alpha'(s) = \alpha(s) \cdot (\text{skew-symmetric matrix}).$$

As we have just seen, this skew-symmetric matrix is just

$$\begin{pmatrix} 0 & -\kappa(s) & 0 \\ \kappa(s) & 0 & -\tau(s) \\ 0 & \tau(s) & 0 \end{pmatrix}.$$

Moreover, this argument shows that skew-symmetry of this matrix is a necessary consequence of the fact that $\mathbf{t}, \mathbf{n}, \mathbf{b}$ are orthonormal. By the same token, we can now present a more illuminating proof that the equations

$$(*) \quad \mathbf{v}_i' = \sum_{j=1}^3 a_{ji} \mathbf{v}_j$$

in the proof of Theorem 11 have an everywhere orthonormal solution. The equation (*) for $\alpha(s) = (\mathbf{v}_1(s), \mathbf{v}_2(s), \mathbf{v}_3(s))$ says that

$$\alpha'(s) = \alpha(s) \cdot a(s).$$

This may simply be regarded as a differential equation *on the manifold* $\text{SO}(3)$, of the type considered in the Addendum to Chapter 5, so its solution is a curve in $\text{SO}(3)$.

The Lie group $\text{SO}(3)$ is playing yet another, hitherto unmentioned, role in Theorem 11. The fact that κ and τ determine c up to a proper Euclidean motion is equivalent to the fact that κ and τ determine $\alpha = (\mathbf{t}, \mathbf{n}, \mathbf{b})$ up to an element of $\text{SO}(3)$, since c is determined up to a translation by \mathbf{t} . In other words, if \bar{c} is another curve with corresponding $\bar{\alpha} = (\bar{\mathbf{t}}, \bar{\mathbf{n}}, \bar{\mathbf{b}})$, and $\bar{\kappa} = \kappa$, $\bar{\tau} = \tau$, then for some $A \in \text{SO}(3)$ we have

$$(\bar{\mathbf{t}}, \bar{\mathbf{n}}, \bar{\mathbf{b}}) = L_A \circ (\mathbf{t}, \mathbf{n}, \mathbf{b}), \quad \text{i.e.,} \quad \bar{\alpha} = L_A \circ \alpha.$$

Now we already have a theorem telling us when the relation $\bar{\alpha} = L_A \circ \alpha$ holds between two maps $\alpha, \bar{\alpha}: [a, b] \rightarrow \text{SO}(3)$. According to Theorem I.10-18, this is the case if and only if $\alpha^*(\omega) = \bar{\alpha}^*(\omega)$ for every left invariant 1-form ω on $\text{SO}(3)$. So κ and τ must have something to do with these left invariant 1-forms on $\text{SO}(3)$. In order to see what is going on here, we begin with a review of some facts about Lie groups.

In Chapter I.10 we defined the natural \mathfrak{g} -valued 1-form ω on a Lie group G by $\omega(a)(\tilde{X}(a)) = X$, where \tilde{X} is the left invariant vector field with $\tilde{X}(e) = X$. Thus ω is the unique left invariant \mathfrak{g} -valued 1-form such that $\omega(e): G_e \rightarrow \mathfrak{g} = G_e$ is the identity. If X_1, \dots, X_n is a basis of \mathfrak{g} , then we can write

$$\omega = \sum_{i=1}^n \omega^i \cdot X_i$$

for certain ordinary left invariant 1-forms ω^i . This equation means that for any tangent vector $Y_a \in G_a$ we have

$$\omega(a)(Y_a) = \sum_{i=1}^n \omega^i(Y_a) \cdot X_i \in \mathfrak{g},$$

the dot denoting multiplication of $X_i \in \mathfrak{g}$ by the real number $\omega^i(a)(Y_a)$. Clearly the ω^i are a basis for the left invariant 1-forms; in fact, the $\omega^i(e)$ are the dual basis to X_1, \dots, X_n . So the natural \mathfrak{g} -valued 1-form ω has all the left invariant 1-forms built into it.

Now for Lie groups G which are subgroups of some $GL(n, \mathbb{R})$ we have an explicit way of finding ω , and hence a basis of left invariant 1-forms. Let P (for “point”) denote the inclusion map of $G \subset GL(n, \mathbb{R}) \subset \mathbb{R}^{n^2}$ into \mathbb{R}^{n^2} . Then $P: G \rightarrow \mathbb{R}^{n^2}$ is an \mathbb{R}^{n^2} -valued function on G . Hence dP is an \mathbb{R}^{n^2} -valued 1-form on G . We can also think of dP as a matrix of ordinary 1-forms on G . This matrix is just

$$dP = (dx^{ij}),$$

except that dx^{ij} here denotes the differential of $x^{ij}|_G$. Notice that dP takes a tangent vector in G_I , i.e., an $n \times n$ matrix M , into itself, so dP corresponds to the identity map of G_I into itself. For any $A \in G$, the map $P \circ L_A: G \rightarrow \mathbb{R}^{n^2}$ is

$$P \circ L_A(B) = A \cdot B,$$

from which it is easy to see that

$$(1) \quad d(P \circ L_A) = A \cdot dP.$$

On G we also have the C^∞ function $B \mapsto B^{-1}$, which we will denote (somewhat confusingly, perhaps) by P^{-1} . For $A \in G$ the map $P^{-1} \circ L_A: G \rightarrow \mathbb{R}^{n^2}$ is

$$(2) \quad P^{-1} \circ L_A(B) = (A \cdot B)^{-1} = B^{-1} \cdot A^{-1} = (P^{-1} \cdot A^{-1})(B).$$

Finally, $P^{-1} \cdot dP$ is a matrix of 1-forms (or an \mathbb{R}^{n^2} -valued 1-form). From (1) and (2) we have

$$\begin{aligned} L_A^*(P^{-1} \cdot dP) &= (P^{-1} \circ L_A) \cdot L_A^*(dP) \\ &= (P^{-1} \circ L_A) \cdot d(L_A^*P) \\ &= (P^{-1} \circ L_A) \cdot d(P \circ L_A) \\ &= (P^{-1} \cdot A^{-1}) \cdot A \cdot dP \\ &= P^{-1} \cdot dP, \end{aligned}$$

so $P^{-1} \cdot dP$ is left invariant. Moreover, for any $M \in G_I$ we have

$$P^{-1} \cdot dP(I)(M) = I^{-1} \cdot dP(I)(M) = M.$$

Hence $P^{-1} \cdot dP$ is the natural \mathfrak{g} -valued 1-form ω on G .

These considerations allow us to determine ω , but it must be remembered that, unless $G = \text{GL}(n, \mathbb{R})$, the forms dx^{ij} are not linearly independent on G . Problem I.10-24 determines ω for $G = \text{GL}(n, \mathbb{R})$, and I.10-23 determines ω for the group $G \subset \text{GL}(2, \mathbb{R})$ consisting of all matrices

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \quad a \neq 0;$$

for this group, x^{11} and x^{12} can be taken as a coordinate system, and ω can be expressed in terms of x^{11} and x^{12} . The more general situation is illustrated by Problem I.10-25, which we shall repeat here. The **special linear group** $\text{SL}(n, \mathbb{R}) \subset \text{GL}(n, \mathbb{R})$ is the set of all matrices of determinant 1. Its Lie algebra, $\mathfrak{sl}(n, \mathbb{R})$, consists of all matrices with trace 0. We can prove this from the fact that

$$\det \exp M = e^{\text{trace } M},$$

in the same way that we found $\mathfrak{o}(n)$; the formula just given is proved in Problem I.10-15. Now we see that, if $x^{11}, x^{12}, x^{21}, x^{22}$ are denoted by x, y, u, v , then for $\text{SL}(2, \mathbb{R})$ we have

$$P^{-1} \cdot dP = \begin{pmatrix} v & -y \\ -u & x \end{pmatrix} \cdot \begin{pmatrix} dx & dy \\ du & dv \end{pmatrix} = \begin{pmatrix} v dx - y du & v dy - y dv \\ -u dx + x du & -u dy + x dv \end{pmatrix}.$$

Remember that x, y, u, v are *not* linearly independent, for the dimension of $\mathfrak{sl}(2, \mathbb{R})$, and hence of $\text{SL}(2, \mathbb{R})$, is clearly 3. In fact, we know that $xv - yu = 1$ on $\text{SL}(2, \mathbb{R})$, which shows that

$$0 = d(xv - yu) = x dv + v dx - y du - u dy$$

on $\text{SL}(2, \mathbb{R})$, i.e., the right side gives 0 when applied to a matrix $M \in \text{SL}(2, \mathbb{R})_A$ for any $A \in \text{SL}(2, \mathbb{R})$. Looking at the formula for $P^{-1} \cdot dP$, this shows that $P^{-1} \cdot dP$ takes such a matrix M to a matrix of trace 0, as it must.

We will now use this circle of ideas to rederive our results about curves, in a more systematic, if less geometric, way. In order to clarify the general nature of the results, we begin once again with curves in the plane, but we now seek the answer to a different classification problem. We already know that if A is a proper Euclidean motion, and $c: [a, b] \rightarrow \mathbb{R}^2$ is any curve, then

- (a) $A \circ c$ is parameterized by arclength whenever c is; we express this fact by saying that arclength is a “natural parameter for curves under the group of proper Euclidean motions”.

- (b) For curves c parameterized by arclength, the curvature function is invariant under proper Euclidean motions, i.e., the function κ for $A \circ c$ equals the function κ for c , whenever A is a proper Euclidean motion. Moreover, the curvature is “a complete set of invariants for curves parameterized by arclength”: if $\bar{\kappa}$ for \bar{c} equals κ for c , then $\bar{c} = A \circ c$ for some proper Euclidean motion A .

We now ask for similar results when A is allowed to be any “special affine motion” [a translation followed by any member of $\text{SL}(2, \mathbb{R})$].

Consider a curve $c: [a, b] \rightarrow \mathbb{R}^2$ for which c' and c'' are always linearly independent. Then $\det(c', c'') \neq 0$ at all points. For simplicity we will assume that $\det(c', c'') > 0$, to avoid writing absolute value signs in various formulas. Our first task is to determine a curve $\alpha_c: [a, b] \rightarrow \text{SL}(2, \mathbb{R})$ analogous to the curve $(\mathbf{t}, \mathbf{n}): [a, b] \rightarrow \text{SO}(2)$. Remember that \mathbf{t} is just c' when c is parameterized by arclength, while \mathbf{n} is basically just chosen so that we will have $(\mathbf{t}, \mathbf{n}) \in \text{SO}(2)$.

For the case of the larger group $\text{SL}(2, \mathbb{R})$ the choice of α_c should be easier. In fact, the choice

$$\alpha_c(t) = (c'(t), c''(t))$$

will work if $\det(c'(t), c''(t)) = 1$. So we ask if there is a reparameterization of c with this property. In other words, is there a function

$$\sigma: [a, b] \rightarrow [0, \ell]$$

such that the reparameterization $\gamma = c \circ \sigma^{-1}$ satisfies

$$\det(\gamma', \gamma'') = 1.$$

Since

$$\begin{aligned} c &= \gamma \circ \sigma \\ c' &= \sigma' \cdot (\gamma' \circ \sigma) \\ c'' &= (\sigma')^2 \cdot (\gamma'' \circ \sigma) + \sigma'' \cdot (\gamma' \circ \sigma), \end{aligned}$$

and thus

$$\begin{aligned} \det(c', c'') &= \det(\sigma' \cdot (\gamma' \circ \sigma), (\sigma')^2 \cdot (\gamma'' \circ \sigma) + \sigma'' \cdot (\gamma' \circ \sigma)) \\ &= (\sigma')^3 \det(\gamma' \circ \sigma, \gamma'' \circ \sigma), \end{aligned}$$

we want

$$\det(c', c'') = (\sigma')^3.$$

We can thus define σ , the “special affine arclength” by

$$\sigma(t) = \int_a^t \sqrt[3]{\det(c'(u), c''(u))} \, du.$$

When c satisfies $\det(c', c'') = 1$, we say that c is “parameterized by special affine arclength”, and the special affine arclength is a natural parameter for curves under the group of special affine motions.

Now consider two curves c, \bar{c} parameterized by special affine arclength. We will have $\bar{c} = A \circ c$ for some special affine motion $A = B \circ \tau$, where $B \in \text{SL}(2, \mathbb{R})$ and τ is a translation, precisely when

$$\alpha_{\bar{c}}(t) = B \cdot \alpha_c(t), \quad \alpha_{\bar{c}} = L_B \circ \alpha_c.$$

By Theorem I.10-18, this is equivalent to the condition

$$\alpha_{\bar{c}}^*(\omega^i) = \alpha_c^*(\omega^i)$$

for every left invariant 1-form ω^i on $\text{SL}(2, \mathbb{R})$, and thus to the condition

$$\alpha_{\bar{c}}^*(\omega) = \alpha_c^*(\omega)$$

for the natural $\mathfrak{sl}(2, \mathbb{R})$ -valued form $\omega = P^{-1} \cdot dP$ on $\text{SL}(2, \mathbb{R})$.

We are thus interested in calculating $\alpha_c^*(P^{-1} \cdot dP)$, which will be a matrix of 1-forms on $[a, b]$

$$\begin{pmatrix} \text{---} dt & \text{---} dt \\ \text{---} dt & \text{---} dt \end{pmatrix};$$

for simplicity we will write all the dt 's after the matrix.

To calculate $\alpha_c^*(P^{-1} \cdot dP)$ we can either use our formula for $P^{-1} \cdot dP$ and write

$$\begin{aligned} \alpha_c^*(P^{-1} \cdot dP) &= \alpha_c^* \begin{pmatrix} v dx - y du & v dy - y dv \\ -u dx + x du & -u dy + x dv \end{pmatrix} \\ &= \begin{pmatrix} v \circ \alpha_c d(x \circ \alpha_c) - \cdots & \cdots \\ \cdots & \cdots \end{pmatrix} \end{aligned}$$

or, what amounts to the same thing, calculate

$$\alpha_c^*(P^{-1} \cdot dP) = \alpha_c^{-1} \cdot d\alpha_c,$$

the entries of which are dt times

$$[\alpha_c(t)]^{-1} \cdot \alpha_c'(t).$$

We have

$$\alpha_c' = \begin{pmatrix} c_1'' & c_1''' \\ c_2'' & c_2''' \end{pmatrix},$$

and using Kramer's rule, together with the fact that $1 = \det \alpha_c = \det(c', c'')$, we find that

$$\alpha_c^{-1} = \begin{pmatrix} c_2'' & -c_1'' \\ -c_2' & c_1' \end{pmatrix}.$$

So

$$\alpha_c^{-1} \cdot \alpha_c' = \begin{pmatrix} c_2'' & -c_1'' \\ -c_2' & c_1' \end{pmatrix} \cdot \begin{pmatrix} c_1'' & c_1''' \\ c_2'' & c_2''' \end{pmatrix}.$$

Since

$$1 = \det(c', c'') = c_1' c_2'' - c_2' c_1'',$$

and thus also

$$0 = c_1' c_2''' - c_2' c_1''',$$

we obtain finally

$$\alpha_c^{-1} \cdot \alpha_c' = \begin{pmatrix} 0 & c_2'' c_1''' - c_1'' c_2''' \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\det(c'', c''') \\ 1 & 0 \end{pmatrix}.$$

Thus, curves parameterized by special affine arclength are determined, up to a special affine motion, by one function, the “special affine curvature”

$$\kappa = \det(c'', c''').$$

It is also possible to give a geometric interpretation of the curvature κ , which we only briefly indicate. We first note that a curve c , parameterized by σ , with constant curvature κ satisfies

$$\begin{aligned} [\alpha_c(\sigma)]^{-1} \cdot \alpha_c'(\sigma) &= \begin{pmatrix} 0 & -\kappa \\ 1 & 0 \end{pmatrix} \\ \alpha_c'(\sigma) &= \alpha_c(\sigma) \cdot \begin{pmatrix} 0 & -\kappa \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

One solution of this matrix differential equation is

$$\alpha_c(\sigma) = \exp \left\{ \sigma \cdot \begin{pmatrix} 0 & -\kappa \\ 1 & 0 \end{pmatrix} \right\},$$

that is,

$$\begin{aligned}\alpha_c(\sigma) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sigma \begin{pmatrix} 0 & -\kappa \\ 1 & 0 \end{pmatrix} + \frac{\sigma^2}{2!} \begin{pmatrix} -\kappa & 0 \\ 0 & -\kappa \end{pmatrix} + \frac{\sigma^3}{3!} \begin{pmatrix} 0 & -\kappa^2 \\ \kappa & 0 \end{pmatrix} \\ &\quad + \frac{\sigma^4}{4!} \begin{pmatrix} \kappa^2 & 0 \\ 0 & \kappa^2 \end{pmatrix} + \cdots \\ &= \begin{pmatrix} \cos \sqrt{\kappa} \sigma & -\sqrt{\kappa} \sin \sqrt{\kappa} \sigma \\ \sqrt{\kappa} \sin \sqrt{\kappa} \sigma & \cos \sqrt{\kappa} \sigma \end{pmatrix}\end{aligned}$$

for $\kappa > 0$, with a similar result involving hyperbolic trigonometric functions for $\kappa < 0$. For $\kappa = 0$ we simply have

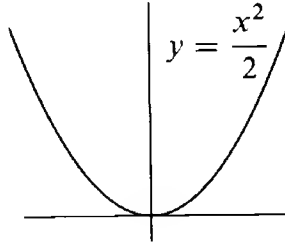
$$\alpha_c(\sigma) = \begin{pmatrix} 1 & 0 \\ \sigma & 1 \end{pmatrix}.$$

The first column of these solutions give

$\kappa = 0$:

$$c'(\sigma) = (1, \sigma)$$

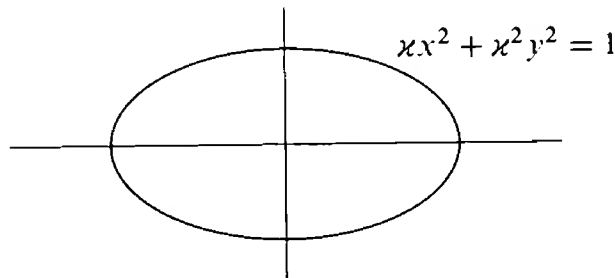
$$c(\sigma) = \text{constant} + (\sigma, \sigma^2/2), \text{ a parabola}$$



$\kappa > 0$:

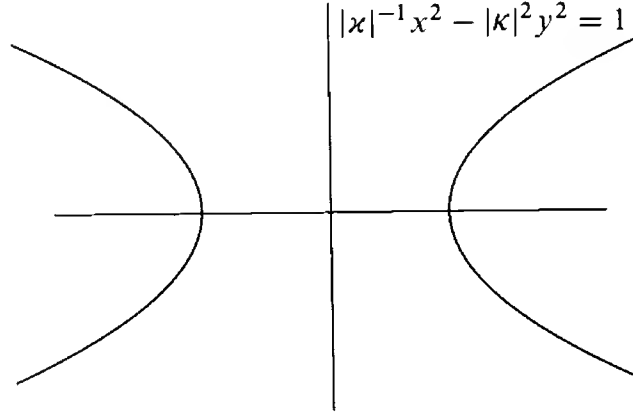
$$c'(\sigma) = (\cos \sqrt{\kappa} \sigma, \sqrt{\kappa} \sin \sqrt{\kappa} \sigma)$$

$$c(\sigma) = \text{constant} + (1/\sqrt{\kappa} \sin \sqrt{\kappa} \sigma, -1/\kappa \cos \sqrt{\kappa} \sigma), \text{ an ellipse}$$



$\kappa < 0$:

$c(\sigma) = \text{constant} + (\sqrt{|\kappa|} \cosh \sqrt{|\kappa|}\sigma, 1/|\kappa| \sinh \sqrt{|\kappa|}\sigma)$, a hyperbola



All other solutions $\alpha_c: \mathbb{R} \rightarrow \text{SL}(2, \mathbb{R})$ are special linear transformations times these, and hence are still conic sections.

Now it turns out that as $\sigma_1, \sigma_2, \sigma_3 \rightarrow \sigma$, the parabola through $c(\sigma_1), c(\sigma_2), c(\sigma_3)$ approaches a given parabola, the **osculating parabola**, whose axis lies in the direction of $c''(s)$. And as $\sigma_1, \sigma_2, \sigma_3, \sigma_4 \rightarrow \sigma$, the conic through the four points $c(\sigma_i)$ approaches a given conic, the **hyperosculating conic**, whose curvature κ is that of c at s . However, we will not prove these facts.

We can now return to curves in space, and apply these ideas to classify curves under the group of proper Euclidean motions. With each curve $c: [a, b] \rightarrow \mathbb{R}^3$ we want to associate a curve $\alpha_c: [a, b] \rightarrow \text{SO}(3)$. Assuming $\det(c', c'', c''') > 0$, the obvious choice is to let $\alpha_c(t)$ be the result of applying the Gram-Schmidt orthonormalization process to these three vectors. Introducing the parameterization by (ordinary) arclength, this means that

$$\alpha_c(s) = \left(c'(s), \frac{c''(s)}{|c''(s)|}, c'(s) \times \frac{c''(s)}{|c''(s)|} \right).$$

We now have

$$\alpha_c'(s) = \alpha_c(s) \cdot \begin{pmatrix} 0 & -k(s) & 0 \\ k(s) & 0 & -t(s) \\ 0 & t(s) & 0 \end{pmatrix};$$

the 0 in position (3, 1) comes about because we chose $c', c''/|c''|$ as the first two columns of α_c , while all other features of the matrix are accounted for by the fact that $\alpha_c^{-1} \cdot \alpha_c' = \alpha_c^*(P^{-1} \cdot dP)$ is skew-symmetric, since $\alpha_c: [a, b] \rightarrow \text{SO}(3)$.

Clearly, curves parameterized by arclength are determined up to proper Euclidean motions by k and t , and these functions are obviously just our old κ and τ . If we had bothered to compute $\alpha_c^{-1} \cdot \alpha_{c'}$ for the original curve, not parameterized by arclength, then the entries in positions (2, 1) and (3, 2) would have given us the formulas for κ and τ , derived earlier, for curves with an arbitrary parameterization.

We can classify curves $c: [a, b] \rightarrow \mathbb{R}^n$ under the group of proper Euclidean motions in a similar way. We assume first of all that $\det(c', \dots, c^{(n)}) > 0$. To obtain a curve $\alpha_c: [a, b] \rightarrow \text{SO}(n)$, we first introduce the parameterization by arclength. The first two columns $\mathbf{v}_1, \mathbf{v}_2$ of α_c will be $c', c''/|c''|$. So the first column of $\alpha_c^{-1} \cdot \alpha_{c'}$, which expresses \mathbf{v}_1' in terms of the \mathbf{v}_j , will be

$$\begin{pmatrix} 0 \\ k_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

for $k_1 = 1/|c''|$. Using skew-symmetry, $\alpha_c^{-1} \cdot \alpha_{c'}$ looks like

$$\alpha_c^{-1} \cdot \alpha_{c'} = \begin{pmatrix} 0 & -k_1 & 0 & \dots & 0 \\ k_1 & 0 & & \times & \\ 0 & & 0 & \times & \\ \vdots & \times & & \ddots & \\ 0 & \times & & & 0 \end{pmatrix}.$$

Since \mathbf{v}_3 , the third column of α_c , is obtained by applying the Gram-Schmidt orthonormalization process to $\mathbf{v}_1, \mathbf{v}_2, c'''$, it is clear that \mathbf{v}_2' will be a linear combination of \mathbf{v}_1 and \mathbf{v}_3 . So the second column of $\alpha_c^{-1} \cdot \alpha_{c'}$ will be

$$\begin{pmatrix} -k_1 \\ 0 \\ k_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

for some function k_2 . Thus $\alpha_c^{-1} \cdot \alpha_c'$ looks like

$$\alpha_c^{-1} \cdot \alpha_c' = \begin{pmatrix} 0 & -k_1 & 0 & \dots & 0 \\ k_1 & 0 & -k_2 & \dots & 0 \\ 0 & k_2 & 0 & & \times \\ \vdots & & \times & \ddots & \times \\ 0 & & & & 0 \end{pmatrix}.$$

Continuing in this way, we find that

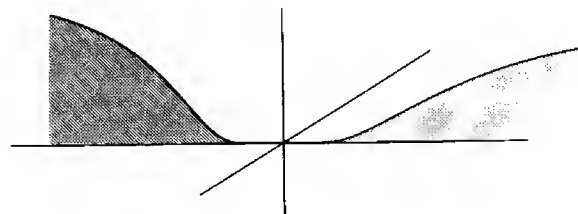
$$\alpha_c^{-1} \cdot \alpha_c' = \begin{pmatrix} 0 & -k_1 & & 0 \\ k_1 & 0 & -k_2 & \\ & k_2 & 0 & \\ & 0 & & \ddots & -k_{n-1} \\ & 0 & & k_{n-1} & 0 \end{pmatrix}$$

for $n - 1$ different “curvature functions” k_1, \dots, k_{n-1} . These $n - 1$ functions classify c up to proper Euclidean motion.

For this classification of curves in \mathbb{R}^n we have had to restrict our attention to curves with $c', \dots, c^{(n)}$ everywhere linearly independent. If we have a curve c such that $c^{(k)}$ is linearly dependent on $c', \dots, c^{(k-1)}$ along a whole interval $[a, b]$, then it is easy to see that on this interval c lies in some $(k - 1)$ -dimensional subspace of \mathbb{R}^n , so that we actually have an easier classification problem on this interval.

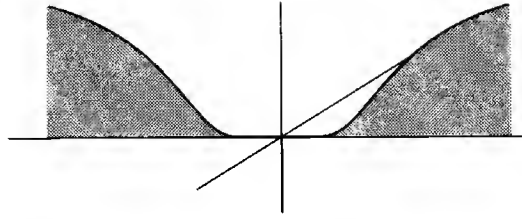
The difficulties arise when we try to piece together the information we obtain for the separate intervals, or if $c^{(k)}$ is linearly dependent on $c', \dots, c^{(k-1)}$ at only isolated points (or at some sequence of points, etc.). For example, how can we hope to distinguish between the curve

$$c(t) = \begin{cases} 0 & t = 0 \\ (t, e^{-1/t^2}, 0) & t > 0 \\ (t, 0, e^{-1/t^2}) & t < 0 \end{cases}$$



and the curve

$$c(t) = \begin{cases} 0 & t = 0 \\ (t, 0, e^{-1/t^2}) & t \neq 0 \end{cases}$$



using functions like curvature and torsion? Of course, for analytic curves these difficulties cannot arise. If any small portion of one analytic curve differs by a proper Euclidean motion from a small portion of a second, then the two analytic curves themselves differ by this proper Euclidean motion. But for C^∞ curves, our restrictions are natural ones to make.

Finally, we will use these ideas to classify curves $c: [a, b] \rightarrow \mathbb{R}^n$ under the group of special affine motions of \mathbb{R}^n . We claim first that there is always a function $\sigma: [a, b] \rightarrow [0, l]$ such that the reparameterization $\gamma = c \circ \sigma^{-1}$ satisfies

$$\det(\gamma', \gamma'', \dots, \gamma^{(n)}) = 1.$$

In fact, since

$$\begin{aligned} c &= \gamma \circ \sigma \\ c' &= \sigma' \cdot (\gamma' \circ \sigma) \\ c'' &= (\sigma')^2 \cdot (\gamma'' \circ \sigma) + () \cdot \gamma' \circ \sigma \\ c''' &= (\sigma')^3 \cdot (\gamma''' \circ \sigma) + () \cdot \gamma'' \circ \sigma + () \cdot \gamma' \circ \sigma \\ &\vdots \end{aligned}$$

we clearly need

$$\det(c', \dots, c^{(n)}) = \sigma' \cdot (\sigma')^2 \cdots (\sigma')^n = (\sigma')^{n(n+1)/2},$$

so we define the “special affine arclength” σ in \mathbb{R}^n by

$$\sigma(t) = \int_a^t \sqrt[n(n+1)/2]{\det(c', \dots, c^{(n)})}.$$

Now if we consider only curves $c: [a, b] \rightarrow \mathbb{R}^n$ with $\det(c', \dots, c^{(n)}) = 1$, then we can define $\alpha_c: [a, b] \rightarrow \text{SL}(n, \mathbb{R})$ by

$$\alpha_c = (c', \dots, c^{(n)}),$$

where each $c^{(i)}$ is considered as a column vector. Solving the equation

$$\alpha_c' = \alpha_c \cdot a,$$

we see that the first column of a must be

$$\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

the second must be

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and so forth. The last column is arbitrary, except that the matrix a must have trace 0, so $\alpha_c^{-1} \cdot \alpha_c'$ must be of the form

$$\alpha_c^{-1} \cdot \alpha_c' = \begin{pmatrix} 0 & 0 & \dots & -\kappa_1 \\ 1 & 0 & & -\kappa_2 \\ 0 & 1 & & \vdots \\ \vdots & \vdots & & -\kappa_{n-1} \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

for $n - 1$ “special affine curvature functions” $\kappa_1, \dots, \kappa_{n-1}$; they determine a curve, parameterized by affine arclength, up to special linear affine maps of \mathbb{R}^n .

The special affine arclength σ in \mathbb{R}^n can be introduced only when

$$\det(c', \dots, c^{(n)}) \neq 0.$$

If along some interval the k vectors $c', \dots, c^{(k)}$ are linearly independent, while the vectors $c', \dots, c^{(k)}, c^{(k+1)}$ are linearly *dependent*, then it is easy to see that along this interval the curve c actually lies in some k -dimensional plane in \mathbb{R}^n , and we can therefore introduce the special affine arclength for \mathbb{R}^k along this interval.

The theory runs into troubles when there are isolated points at which

$$\det(c', \dots, c^{(n)}) = 0.$$

In this respect it is similar to the ordinary theory of curves, but the situation is still different, because in the special affine case the function σ which we choose as the special affine arclength will actually depend on the number k for which $c', \dots, c^{(k)}$ are linearly independent—for curves in higher dimensional Euclidean spaces the special affine arclength involves more derivatives. The special affine curvature functions $\kappa_1, \kappa_2, \kappa_3, \dots$ also involve higher derivatives than the corresponding curvature functions $\kappa_1, \kappa_2, \kappa_3, \dots$. This is just what we ought to expect—since the group $\mathrm{SL}(n, \mathbb{R})$ is bigger than $\mathrm{SO}(n)$, we have to build more complicated things from our curve c before we can find something which is invariant under $\mathrm{SL}(n, \mathbb{R})$.

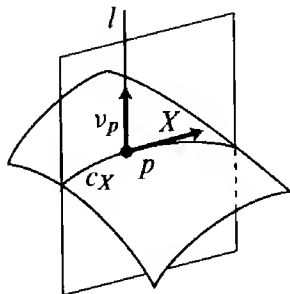
CHAPTER 2

WHAT THEY KNEW ABOUT SURFACES BEFORE GAUSS

Having traversed nearly the whole history of the study of curves, in one chapter, we now turn back to the beginnings of surface theory, and start the long journey toward the modern theory of higher dimensional manifolds. Even though the study of surfaces had begun long before the Serret-Frenet formulas appeared in 1847, the theory of plane curves, at least, was well understood.

The initial study of surfaces in \mathbb{R}^3 began in a way that seems natural enough. Since we have a theory of curves in the plane, we may hope to describe surfaces by investigating the curves in which the surface intersects various planes. Indeed it turns out that we can describe the curvature of such curves in a surprisingly nice and precise way.

The first results along this line, due to Euler, date from 1760. Through a point p on a surface $M \subset \mathbb{R}^3$ we construct the line l which is perpendicular to M_p . For each unit vector $X \in M_p$ we can then consider the plane through p



which contains both X and l . The intersection of this plane and M is the image of a curve c_X with $c_X(0) = p$; we will also suppose c parameterized by arclength, so that $c_X'(0) = X$. We orient all these planes through l by choosing a vector v_p perpendicular to M_p and orienting the plane through X and l so that X, v_p is positively oriented. Then c_X has a signed curvature at 0, which will be denoted by κ_X .

When we replace X by $-X$, the curve c_{-X} is just c_X traversed in the opposite direction. Since the plane through X and v_p now receives the opposite orientation, the curvature κ_{-X} equals the curvature κ_X . Thus, $X \mapsto \kappa_X$ may be thought of as a function of directions in M_p . Euler discovered a striking fact about the curvatures in these different directions:

1. THEOREM (EULER). If the κ_X are not all equal, then there is precisely one direction, represented by a unit vector X_1 , say, in which κ_X has a minimum value $k_1 = \kappa_{X_1}$, and one in which it has a maximum value $k_2 = \kappa_{X_2}$. These two directions are *perpendicular*, and if X makes an angle of θ with X_1 , then

$$\kappa_X = k_1 \cos^2 \theta + k_2 \sin^2 \theta.$$

(Notice that we can just as well write $\kappa_X = k_2 \cos^2 \phi + k_1 \sin^2 \phi$, where ϕ is the angle which X makes with X_2 , since $\phi = \pi/2 - \theta$, and consequently $\cos^2 \theta = \sin^2 \phi$ and $\sin^2 \theta = \cos^2 \phi$. There is also a certain ambiguity in the statement of Euler's theorem which does not affect the final result: If we change the vector v_p so that it points in the opposite direction, then *all* curvatures κ_X are changed to their negatives.)

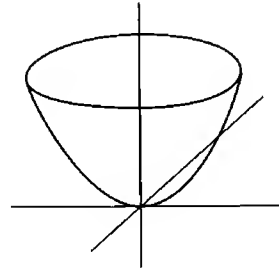
I do not know Euler's proof of his theorem, and for reasons that will appear later, I am sure that the proof to be presented here is much simpler than the original. Nevertheless, it is sufficiently classical in spirit to serve as an historical substitute for Euler's.

PROOF. We begin by choosing our coordinate system so that $p = (0, 0, 0)$ and so that the tangent plane at p is the (x, y) -plane, which means that in a neighborhood of p the surface M is $\{(x, y, z) : z = f(x, y)\}$, where

$$f(0, 0) = 0$$

$$\frac{\partial f}{\partial x}(0, 0) = 0$$

$$\frac{\partial f}{\partial y}(0, 0) = 0.$$



We now maintain that by rotating the (x, y) -plane we can arrange to have

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = 0.$$

To see this, we first recall that rotation through an angle of θ radians is the linear transformation R_θ with matrix $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$. If we rotate the (x, y) -plane by an angle of θ , then M becomes the graph of

$$\begin{aligned} f_\theta &= f \circ R_{-\theta}, \\ f_\theta(x, y) &= f(x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta). \end{aligned}$$

So

$$\frac{\partial f_\theta}{\partial y}(x, y) = D_1 f(x, y)(-\sin \theta) + D_2 f(x, y)(\cos \theta)$$

$$\begin{aligned} \frac{\partial^2 f_\theta}{\partial x \partial y}(x, y) &= D_{11} f(x, y)(-\sin \theta \cos \theta) + D_{12} f(x, y)[-\sin^2 \theta + \cos^2 \theta] \\ &\quad + D_{22} f(x, y)(\sin \theta \cos \theta) \end{aligned}$$

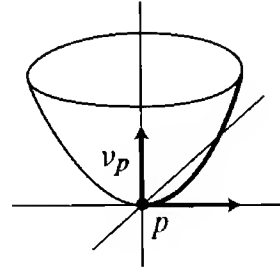
$$\frac{\partial^2 f_\theta}{\partial x \partial y}(0, 0) = (\cos 2\theta) D_{12} f(0, 0) + \frac{D_{22} f(0, 0) - D_{11} f(0, 0)}{2} \sin 2\theta.$$

In order to have $\partial^2 f_\theta / \partial x \partial y(0, 0) = 0$, we just choose θ so that

$$\begin{aligned} \tan 2\theta &= \frac{2D_{12} f(0, 0)}{D_{11} f(0, 0) - D_{22} f(0, 0)} && \text{if } D_{11} f(0, 0) \neq D_{22} f(0, 0) \\ \theta &= \pi/4 && \text{if } D_{11} f(0, 0) = D_{22} f(0, 0). \end{aligned}$$

Having made this choice of coordinates, we now look at the various planes containing l , i.e., containing the z -axis. First, the (x, z) -plane intersects the surface in the curve

$$c(t) = (t, f(t, 0)).$$



Its curvature at 0 is therefore

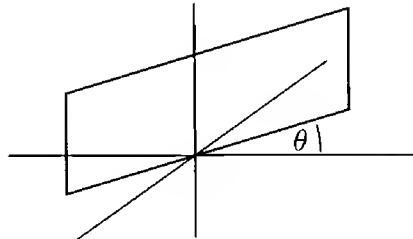
$$k_1 = \frac{\dot{c}_1 \ddot{c}_2 - \dot{c}_2 \ddot{c}_1}{(\dot{c}_1^2 + \dot{c}_2^2)^{3/2}} = \frac{\partial^2 f}{\partial x^2}(0, 0).$$

This result holds when we give the (x, z) -plane its usual orientation, which is the same as the orientation making $(e_1)_p, v_p$ positively oriented when we choose v_p to be $(e_3)_p$. Similarly, the (y, z) -plane intersects the surface in the curve $c(t) = (0, t, f(0, t))$ with curvature

$$k_2 = \frac{\partial^2 f}{\partial y^2}(0, 0).$$

Finally, the plane through the z -axis which makes an angle of θ with the x -axis intersects the surface in the curve

$$c(t) = (t, f(t \cos \theta, t \sin \theta)),$$

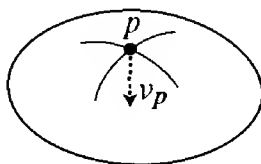


with curvature

$$\begin{aligned}\kappa &= \frac{d^2}{dt^2} f(t \cos \theta, t \sin \theta) = \cos^2 \theta \frac{\partial^2 f}{\partial x^2}(0, 0) + \sin^2 \theta \frac{\partial^2 f}{\partial y^2}(0, 0) \\ &= k_1 \cos^2 \theta + k_2 \sin^2 \theta.\end{aligned}$$

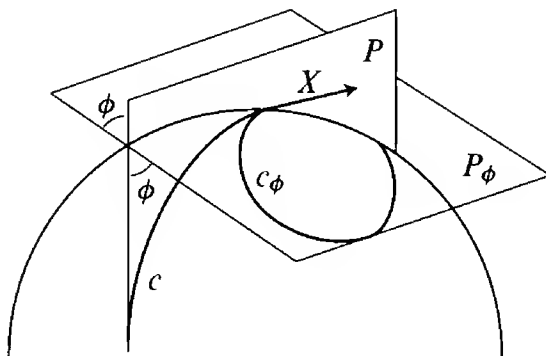
This formula shows that κ always lies between k_1 and k_2 , the curvatures in two perpendicular directions, and thus proves the whole theorem. ♦

Notice that if $\partial^2 f / \partial x^2, \partial^2 f / \partial y^2 > 0$, so that the surface lies above the (x, y) -plane near p , then our choice of v_p makes $k_1, k_2 > 0$. In general, if a surface locally lies on one side of its tangent plane through p , then the choice of v_p as a vector pointing toward this side makes $k_1, k_2 > 0$. If our surface is the boundary of a convex set in \mathbb{R}^3 , we must therefore choose the *inward* pointing normal to obtain positive curvatures. If k_1 and k_2 are of different signs, there



is generally no such way to distinguish a direction for v_p .

The other main result about curves on surfaces, due to Meusnier, came shortly afterwards, in 1776. Meusnier completed Euler's investigations by finding the curvatures of the curves obtained by intersecting the surface M with *any* plane through $p \in M$. If P is the plane through p which contains l and a unit vector $X \in M_p$, then any other plane P_ϕ which contains X can be described by giving the angle ϕ which it makes with P . Let c be the intersection of P and M , with

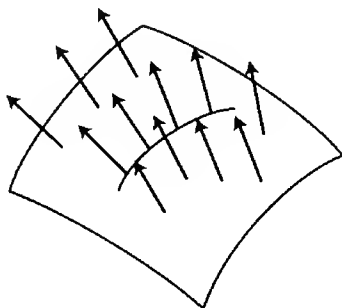


$c(0) = p$ and c parameterized by arclength, so that $c'(0) = X$, and let c_ϕ be the corresponding intersection of P_ϕ and M . Meusnier's theorem describes the curvature κ_ϕ of c_ϕ at 0 in terms of the curvature κ_X of c at 0:

$$\kappa_\phi \cdot \cos \phi = \kappa_X.$$

For the case of the unit sphere, pictured above, the curve c is a great circle, of radius 1, while c_ϕ is a circle of radius $\cos \phi$, so that $\kappa_\phi = 1/\cos \phi$. Naturally, in Meusnier's theorem we must restrict ϕ to be less than $\pi/2$. At this angle, the plane P_ϕ is just the tangent plane, and does not generally intersect M in a curve at all; for ϕ close to $\pi/2$, the plane P_ϕ intersects M in a curve of very large curvature (if $\kappa_X \neq 0$).

For Meusnier's theorem we supply a proof which is decidedly non-classical in spirit, but which will be useful to have later on. We first define a function v in a neighborhood $U \subset M$ of p such that $v(q) \in \mathbb{R}^3$ is a unit vector and $v(q)_q \in \mathbb{R}^3_q$ is perpendicular to M_q for all $q \in U$. There are two choices for each $v(q)$; in order to obtain a continuous function $v: U \rightarrow M$, we orient U , and then pick $v(q)$ so that $v(q), v, w$ is positively oriented for v_q, w_q positively oriented in M_q .



For any curve c in M with $c(0) = p$ we have

$$\langle c'(s), v(c(s)) \rangle = 0 \quad \text{for all } s.$$

Differentiating this equation, we have

$$\langle c''(0), v(p) \rangle = - \left\langle c'(0), \frac{dv(c(s))}{ds} \Big|_{s=0} \right\rangle.$$

Now the vector $dv(c(s))/ds|_{s=0}$, with components $dv^i(c(s))/ds|_{s=0}$, depends only on $c'(0) = X$; in fact, it equals $(X(v^1), X(v^2), X(v^3))$. To see this, just remember that to operate on a function $f: M \rightarrow \mathbb{R}$ with a tangent vector $X_p \in M_p$, we can take any curve c with $c'(0) = X_p$ and then $X_p(f) = df(c(t))/dt|_{t=0}$. We can thus write

$$(*) \quad \langle c''(0), v(p) \rangle = a(X) \quad X = c'(0).$$

Meusnier's theorem follows directly from this equation:

2. THEOREM (MEUSNIER). Let P be the plane through $\nu(p)$ and $X \in M_p$, and let κ_ϕ be the curvature of the curve on M cut out by a plane P_ϕ containing X and making an angle of ϕ with the plane P . Then

$$\kappa_\phi \cdot \cos \phi = \kappa_X.$$

PROOF. First apply (*) to the curve c cut out by P . The second derivative $c''(0)$ of this curve is in P , and is also perpendicular to $c'(0) = X$. Consequently, it is a multiple of $\nu(p)$. Since P has been oriented so that $X, \nu(p)$ is positively oriented, it follows that $c''(0) = \kappa_X \cdot \nu(p)$, so

$$a(X) = \langle c''(0), \nu(p) \rangle = \kappa_X.$$

On the other hand, since $c_\phi'(0) = c'(0) = X$, equation (*) also gives

$$(1) \quad \kappa_X = a(X) = \langle c_\phi''(0), \nu(p) \rangle.$$

We can write $c_\phi''(0) = \kappa_\phi \cdot \nu_\phi$ where ν_ϕ is a unit vector which lies in P_ϕ and is perpendicular to X . Then ν_ϕ makes an angle of ϕ with $\nu(p)$, which means that

$$(2) \quad \langle \nu_\phi, \nu(p) \rangle = \cos \phi.$$

Combining equations (1) and (2), we obtain

$$\begin{aligned} \kappa_X &= \langle c_\phi''(0), \nu(p) \rangle \\ &= \langle \kappa_\phi \cdot \nu_\phi, \nu(p) \rangle \\ &= \kappa_\phi \cdot \cos \phi. \quad \blacklozenge \end{aligned}$$

Despite the appealing simplicity of these results, there is something dissatisfying about this whole approach of dissecting a surface into curves; we never seem to really get our hands on the surface itself. To do this, we must move forward 50 years in time.

CHAPTER 3

THE CURVATURE OF SURFACES IN SPACE

A. HOW TO READ GAUSS

The single most important work in the history of differential geometry is Karl Friedrich Gauss' paper, in Latin, of 1827: *Disquisitiones generales circa superficies curvas*. The following translation of (part of) this paper is basically the one published* by The Princeton University Library, 1902, except that it adheres even more closely to the notation and typographic disposition of the original.

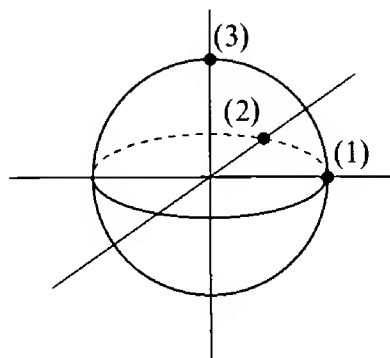
In addition, it has been supplemented with remarks designed to make this first confrontation with classical differential geometry much less painful. The translation of Gauss' paper appears to the right—on odd-numbered pages—while corresponding remarks appear to the left.

Although Part B of this chapter is an exposition of Gauss' results, in modern notation, a preliminary reading of Gauss' great work is heartily recommended; and since many of the difficulties will be clarified in Part B, as a general rule it is a good idea to read on, even if a particular section makes very little sense!

*A reprinting was produced in 1965 by Raven Press, but this is also out of print.

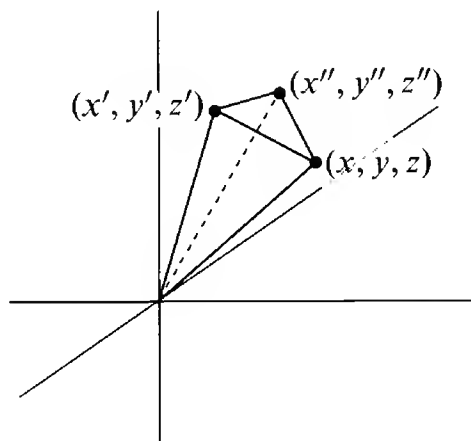
REMARKS ON GAUSS' PAPER

§1. Notice that (1), (2), (3) are used as the names of certain points [(1) = (1, 0, 0), etc.], a circumstance that is easy to forget later on.



§2. This section gives a complicated proof, using spherical trigonometry, that the volume of the pyramid shown below is

$$\frac{1}{6} \left| \det \begin{pmatrix} x & y & z \\ x' & y' & z' \\ x'' & y'' & z'' \end{pmatrix} \right|,$$



and also includes remarks about the significance of the sign of the determinant. This result is equivalent to the well-known fact that $|\det A|$ is the volume of the parallelepiped spanned by the rows of A .

Almost all of this section can simply be skipped, except for noting that the

GENERAL INVESTIGATIONS
OF
CURVED SURFACES

1.

Investigations, in which the directions of various straight lines in space are to be considered, attain a high degree of clearness and simplicity if we employ, as an auxiliary, a sphere of radius = 1 described about an arbitrary center, and suppose the different points of the sphere to represent the directions of straight lines parallel to the radii ending at these points. As the position of every point in space is determined by three coordinates, that is to say, the distances of the point from three mutually perpendicular fixed planes, it is necessary to consider, first of all, the directions of the axes perpendicular to these planes. The points on the sphere, which represent these directions, we shall denote by (1), (2), (3). The distance of any one of these points from either of the other two will be a quadrant; and we shall suppose that the directions of the axes are those in which the corresponding coordinates increase.

2.

It will be advantageous to bring together here some propositions which are frequently used in questions of this kind.

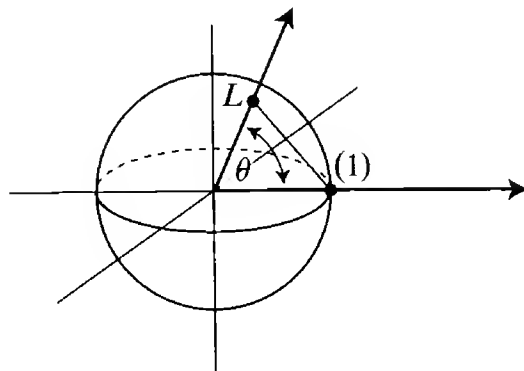
I. The angle between two intersecting straight lines is measured by the arc between the points on the sphere which correspond to the directions of the lines.

II. The orientation of any plane whatever can be represented by the great circle on the sphere, the plane of which is parallel to the given plane.

III. The angle between two planes is equal to the spherical angle between the great circles representing them, and, consequently, is also measured by the arc intercepted between the poles of these great circles. And, in like manner, the angle of inclination of a straight line to a plane is measured by the arc drawn from the point which corresponds to the direction of the line, perpendicular to the great circle which represents the orientation of the plane.

IV. Letting $x, y, z; x', y', z'$ denote the coordinates of two points, r the distance between them, and L the point on the sphere which represents the

expression $\cos (1)L$, which will also appear later on, means the cosine of the angle θ between the ray from $(0,0,0)$ through the point $L = (a,b,c)$ on the



sphere and the ray from $(0,0,0)$ through $(1) = (1,0,0)$. Thus, for the usual inner product $\langle \ , \ \rangle$ on \mathbb{R}^3 we have

$$a = \langle L, (1) \rangle = 1 \cdot 1 \cdot \cos \theta,$$

so $\cos (1)L$ is the first component of L , and similarly for $\cos (2)L$ and $\cos (3)L$. [The original contains $\cos (1)L^2$ instead of $\cos^2 (1)L$, etc., and multiplication is always indicated with a low dot \cdot rather than a centered dot.]

direction of the line drawn from the first point to the second, we shall have

$$x' = x + r \cos (1)L, \quad y' = y + r \cos (2)L, \quad z' = z + r \cos (3)L$$

V. From this it follows at once that, generally,

$$\cos^2 (1)L + \cos^2 (2)L + \cos^2 (3)L = 1$$

and also, if L' denote any other point on the sphere,

$$\cos (1)L \cdot \cos (1)L' + \cos (2)L \cdot \cos (2)L' + \cos (3)L \cdot \cos (3)L' = \cos LL'$$

VI. THEOREM. If L, L', L'', L''' denote four points on the sphere, and A the angle which the arcs $LL', L''L'''$ make at their point of intersection, then we shall have

$$\cos LL'' \cdot \cos L'L''' - \cos LL''' \cdot \cos L'L'' = \sin LL' \cdot \sin L''L''' \cdot \cos A$$

Demonstration. Let A denote also the point of intersection itself, and set

$$AL = t, \quad AL' = t', \quad AL'' = t'', \quad AL''' = t'''$$

Then we shall have

$$\begin{aligned} \cos LL'' &= \cos t \cos t'' + \sin t \sin t'' \cos A \\ \cos L'L''' &= \cos t' \cos t''' + \sin t' \sin t''' \cos A \\ \cos LL''' &= \cos t \cos t''' + \sin t \sin t''' \cos A \\ \cos L'L'' &= \cos t' \cos t'' + \sin t' \sin t'' \cos A \end{aligned}$$

and consequently,

$$\begin{aligned} &\cos LL'' \cdot \cos L'L''' - \cos LL''' \cdot \cos L'L'' \\ &= \cos A (\cos t \cos t'' \sin t' \sin t''' + \cos t' \cos t''' \sin t \sin t'' \\ &\quad - \cos t \cos t''' \sin t' \sin t'' - \cos t' \cos t'' \sin t \sin t''') \\ &= \cos A (\cos t \sin t' - \sin t \cos t') (\cos t'' \sin t''' - \sin t'' \cos t''') \\ &= \cos A \cdot \sin (t' - t) \cdot \sin (t''' - t'') \\ &= \cos A \cdot \sin LL' \cdot \sin L''L''' \end{aligned}$$

But as there are for each great circle two branches going out from the point A , these two branches form at this point two angles whose sum is 180° . But our analysis shows that those branches are to be taken whose directions are in the

sense from the point L to L' , and from the point L'' to L''' ; and since great circles intersect in two points, it is clear that either of the two points can be chosen arbitrarily. Also, instead of the angle A , we can take the arc between the poles of the great circles of which the arcs LL' , $L''L'''$ are parts. But it is evident that those poles are to be chosen which are similarly placed with respect to these arcs; that is to say, when we go from L to L' and from L'' to L''' , both of the two poles are to be on the right, or both on the left.

VII. Let L, L', L'' be three points on the sphere and set, for brevity,

$$\begin{aligned}\cos(1)L &= x, & \cos(2)L &= y, & \cos(3)L &= z \\ \cos(1)L' &= x', & \cos(2)L' &= y', & \cos(3)L' &= z' \\ \cos(1)L'' &= x'', & \cos(2)L'' &= y'', & \cos(3)L'' &= z''\end{aligned}$$

and also

$$xy'z'' + x'y''z + x''yz' - xy''z' - x'y'z'' - x''y'z = \Delta$$

Let λ denote the pole of the great circle of which LL' is a part, this pole being the one that is placed in the same position with respect to this arc as the point (1) is with respect to the arc (2)(3). Then we shall have, by the preceding theorem, $yz' - y'z = \cos(1)\lambda \cdot \sin(2)(3) \cdot \sin LL'$, or, because $(2)(3) = 90^\circ$,

$$\begin{aligned}yz' - y'z &= \cos(1)\lambda \cdot \sin LL', & \text{and similarly} \\ zx' - z'x &= \cos(2)\lambda \cdot \sin LL' \\ xy' - x'y &= \cos(3)\lambda \cdot \sin LL'\end{aligned}$$

Multiplying these equations by x'', y'', z'' respectively, and adding, we obtain, by means of the second of the theorems deduced in V,

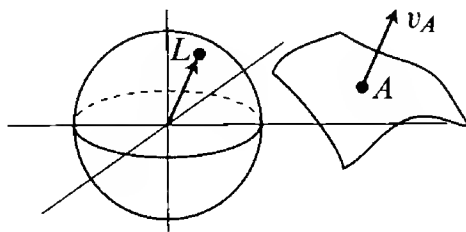
$$\Delta = \cos \lambda L'' \cdot \sin LL'$$

Now there are three cases to be distinguished. *First*, when L'' lies on the great circle of which the arc LL' is a part, we shall have $\lambda L'' = 90^\circ$, and consequently, $\Delta = 0$. If L'' does not lie on that great circle, the *second* case will be when L'' is on the same side as λ ; the *third* case when they are on opposite sides. In the last two cases the points L, L', L'' will form a spherical triangle, and in the second case these points will lie in the same order as the points (1), (2), (3), and in the opposite order in the third case. Denoting the angles of this triangle simply by L, L', L'' and the perpendicular drawn on the sphere from the point L'' to the side LL' by p , we shall have

$$\sin p = \sin L \cdot \sin LL'' = \sin L' \cdot \sin L'L'', \quad \text{and} \quad \lambda L'' = 90^\circ \mp p$$

§3. This section merely defines (or tries to define) a differentiable surface, and its tangent plane at a point.

§4. At a point $A = (x, y, z)$ in the surface we have a unit normal vector v_A , and $v \in S^2 \subset \mathbb{R}^3$ is what Gauss calls L . The expression $\cos(1)L$ means the cosine of



the angle between the rays from $(0, 0, 0)$ through L and through $(1) = (1, 0, 0)$ (c.f. page 58). So X, Y, Z are just the components of L . Thus X, Y, Z can be considered as functions on the surface [$X(A) =$ first component of v , for v_A a unit normal at A , etc.].

Gauss now nonchalantly introduces infinitely small quantities. The goal of his

the upper sign being taken for the second case, the lower for the third. From this it follows that

$$\pm\Delta = \sin L \cdot \sin LL' \cdot \sin LL'' = \sin L' \cdot \sin LL' \cdot \sin L'L'' = \sin L'' \cdot \sin LL'' \cdot \sin L'L''$$

Moreover, it is evident that the first case can be regarded as contained in the second or third, and it is easily seen that the expression $\pm\Delta$ represents six times the volume of the pyramid formed by the points L, L', L'' and the center of the sphere. Whence, finally, it is clear that the expression $\pm\frac{1}{6}\Delta$ expresses generally the volume of any pyramid contained between the origin of coordinates and the three points whose coordinates are $x, y, z; x', y', z'; x'', y'', z''$.

3.

A curved surface is said to possess continuous curvature at one of its points A , if the directions of all the straight lines drawn from A to points of the surface at an infinitely small distance from A are deflected infinitely little from one and the same plane passing through A . This plane is said to *touch* the surface at the point A . If this condition is not satisfied for any point, the continuity of the curvature is here interrupted, as happens, for example, at the vertex of a cone. The following investigations will be restricted to such surfaces, or to such parts of surfaces, as have the continuity of their curvature nowhere interrupted. We shall only observe now that the methods used to determine the position of the tangent plane lose their meaning at singular points, in which the continuity of the curvature is interrupted, and must lead to indeterminate solutions.

4.

The orientation of the tangent plane is most conveniently studied by means of the direction of the straight line normal to the plane at the point A , which is also called the normal to the curved surface at the point A . We shall represent the direction of this normal by the point L on the auxiliary sphere, and we shall set

$$\cos(1)L = X, \quad \cos(2)L = Y, \quad \cos(3)L = Z$$

and denote the coordinates of the point A by x, y, z . Also let $x + dx, y + dy, z + dz$ be the coordinates of another point A' on the curved surface; ds its distance from A , which is infinitely small; and finally, let λ be the point on the sphere representing the direction of the element AA' . Then we shall have

$$dx = ds \cdot \cos(1)\lambda, \quad dy = ds \cdot \cos(2)\lambda, \quad dz = ds \cdot \cos(3)\lambda$$

initial manipulations is the equation

$$X dx + Y dy + Z dz = 0.$$

If x, y, z are considered as functions on the surface (that is, as the restriction to the surface of the standard coordinate functions on \mathbb{R}^3), then this equation is literally true, interpreting dx, dy, dz as modern differentials. It should be easy to see this (remember how X, Y, Z are defined). Also try to follow Gauss' argument.

The rest of section 4 gives formulas for X, Y, Z in terms of different descriptions of the surface; in each case the formulas are paired with their negatives, since there are two different choices for the unit normal vector:

(1) If the surface is $\{p \in \mathbb{R}^3 : W(p) = 0\}$, for $W: \mathbb{R}^3 \rightarrow \mathbb{R}$, then

$$X = \frac{P}{\sqrt{P^2 + Q^2 + R^2}}, \quad \text{where } P = D_1 W, Q = D_2 W, R = D_3 W, \text{ etc.}$$

[The original has $XX + YY + ZZ = 1$, and $PP + QQ + RR$ for $P^2 + Q^2 + R^2$, and so forth, with a superscript ² used only for the square of a term that is not a single letter.]

(2) If the surface is the image of $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ [Gauss writes dx for $d(x \circ f) = df^1$, etc.] and

$$\begin{aligned} a &= D_1 f^1, & a' &= D_2 f^1 \\ b &= D_1 f^2, & b' &= D_2 f^2 \\ c &= D_1 f^3, & c' &= D_2 f^3 \end{aligned}$$

and, since λL must be equal to 90° ,

$$X \cos (1)\lambda + Y \cos (2)\lambda + Z \cos (3)\lambda = 0$$

By combining these equations we obtain

$$X dx + Y dy + Z dz = 0$$

There are two general methods for defining the nature of a curved surface. The *first* uses the equation between the coordinates x, y, z , which we may suppose reduced to the form $W = 0$, where W will be a function of the indeterminants x, y, z . Let the complete differential of the function W be

$$dW = P dx + Q dy + R dz$$

and on the curved surface we shall have

$$P dx + Q dy + R dz = 0$$

and consequently,

$$P \cos (1)\lambda + Q \cos (2)\lambda + R \cos (3)\lambda = 0$$

Since this equation, as well as the one we have established above, must be true for the directions of all elements ds on the curved surface, we easily see that X, Y, Z must be proportional to P, Q, R respectively, and consequently, since

$$X^2 + Y^2 + Z^2 = 1$$

we shall have either

$$X = \frac{P}{\sqrt{(P^2 + Q^2 + R^2)}}, \quad Y = \frac{Q}{\sqrt{(P^2 + Q^2 + R^2)}}, \quad Z = \frac{R}{\sqrt{(P^2 + Q^2 + R^2)}}$$

or

$$X = \frac{-P}{\sqrt{(P^2 + Q^2 + R^2)}}, \quad Y = \frac{-Q}{\sqrt{(P^2 + Q^2 + R^2)}}, \quad Z = \frac{-R}{\sqrt{(P^2 + Q^2 + R^2)}}$$

The *second* method expresses the coordinates in the form of functions of two variables, p, q . Suppose that differentiation of these functions gives

$$dx = a dp + a' dq$$

$$dy = b dp + b' dq$$

$$dz = c dp + c' dq$$

then

$$X = \frac{bc' - cb'}{\Delta}, \text{ etc.}$$

(3) If the surface is $\{(x, y, z) : z = f(x, y)\}$ for $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$, then

$$X = \frac{t}{\sqrt{1 + t^2 + u^2}}, \quad \text{where } t = D_1 f, u = D_2 f, \text{ etc.}$$

It should not be hard to work out these results, using our terminology. Again, it is instructive to follow Gauss' derivations as well.

§5. This section talks about orienting the surface, so that one can choose between the two unit normals.

Substituting these values in the formula given above, we obtain

$$(aX + bY + cZ) dp + (a'X + b'Y + c'Z) dq = 0$$

Since this equation must hold independently of the values of the differentials dp , dq , we evidently shall have

$$aX + bY + cZ = 0, \quad a'X + b'Y + c'Z = 0$$

From this we see that X, Y, Z will be proportional to the quantities

$$bc' - cb', \quad ca' - ac', \quad ab' - ba'$$

Hence, on setting, for brevity,

$$\sqrt{((bc' - cb')^2 + (ca' - ac')^2 + (ab' - ba')^2)} = \Delta$$

we shall have either

$$X = \frac{bc' - cb'}{\Delta}, \quad Y = \frac{ca' - ac'}{\Delta}, \quad Z = \frac{ab' - ba'}{\Delta}$$

or

$$X = \frac{cb' - bc'}{\Delta}, \quad Y = \frac{ac' - ca'}{\Delta}, \quad Z = \frac{ba' - ab'}{\Delta}$$

With these two general methods is associated a *third*, in which one of the coordinates, z , say, is expressed in the form of a function of the other two, x, y . This method is evidently only a particular case either of the first method, or of the second. If we set

$$dz = t dx + u dy$$

we shall have either

$$X = \frac{-t}{\sqrt{(1+t^2+u^2)}}, \quad Y = \frac{-u}{\sqrt{(1+t^2+u^2)}}, \quad Z = \frac{1}{\sqrt{(1+t^2+u^2)}}$$

or

$$X = \frac{t}{\sqrt{(1+t^2+u^2)}}, \quad Y = \frac{u}{\sqrt{(1+t^2+u^2)}}, \quad Z = \frac{-1}{\sqrt{(1+t^2+u^2)}}$$

5.

The two solutions found in the preceding article evidently refer to opposite points of the sphere, or to opposite directions, as one would expect, since the

§6. In this section Gauss considers the map ν , from the surface to S^2 , which takes A to the unit vector ν which is normal to the surface at that point. The map ν can be used to take any subset R of the surface to a subset $\nu(R)$ of S^2 . The area of $\nu(R)$ is referred to by Gauss as the *total curvature* of R . Then the

normal may be drawn toward either of the two sides of the curved surface. If we wish to distinguish between the two regions bordering upon the surface, and call one the exterior region and the other the interior region, we can then assign to each of the two normals its appropriate solution by aid of the theorem derived in Art. 2 (VII), and at the same time establish a criterion for distinguishing the one region from the other.

In the first method, such a criterion is to be drawn from the sign of the quantity W . Indeed, generally speaking, the curved surface divides those regions of space in which W keeps a positive value from those in which the value of W becomes negative. In fact, it is easily seen from this theorem that, if W takes a positive value toward the exterior region, and if the normal is supposed to be drawn outwardly, the first solution is to be taken. Moreover, it will be easy to decide in any case whether the same rule for the sign of W is to hold throughout the entire surface, or whether for different parts there will be different rules. As long as the coefficients P, Q, R have finite values and do not all vanish at the same time, the law of continuity will prevent any change.

If we follow the second method, we can imagine two systems of curved lines on the curved surface, one system for which p is variable, q constant; the other for which q is variable, p constant. The respective positions of these lines with reference to the exterior region will decide which of the two solutions must be taken. In fact, whenever the three lines, namely, the branch of the line of the former system going out from the point A as p increases, the branch of the line of the latter system going out from the point A as q increases, and the normal drawn toward the exterior region, are *similarly* placed as the x, y, z axes respectively from the origin of abscissas (e.g., if, both for the former three lines and for the latter three, we can conceive the first directed to the left, the second to the right, and the third upward), the first solution is to be taken. But whenever the relative position of the three lines is opposite to the relative position of the x, y, z axes, the second solution will hold.

In the third method, it is to be seen whether, when z receives a positive increment, x and y remaining constant, the point crosses toward the exterior or the interior region. In the former case, for the normal drawn outward, the first solution holds; in the latter case, the second.

6.

Just as each definite point on the curved surface is made to correspond to a definite point on the sphere, by the direction of the normal to the curved surface which is transferred to the surface of the sphere, so also any line whatever, or any figure whatever, on the latter will be represented by a corresponding line

curvature at a point A in the surface is defined as

$$\frac{\text{total curvature of } R}{\text{area of } R}$$

where R is the “surface element” at A , which is supposed to have infinitely small area. As a first approximation to what Gauss is trying to say, we might define the curvature as

$$\lim \frac{\text{total curvature of } R}{\text{area of } R}$$

where the limit is taken as R approaches the point A . It is not *a priori* so clear whether this limit exists, or if it depends on the way in which R “approaches” A .

Gauss also gives considerable discussion to the sign of the curvature.

or figure on the former. In the comparison of two figures corresponding to one another in this way, one of which will be as the map of the other, two important points are to be considered, one when quantity alone is considered, the other when, disregarding quantitative relations, position alone is considered.

The first of these important points will be the basis of some ideas which it seems judicious to introduce into the theory of curved surfaces. Thus, to each part of a curved surface inclosed within definite limits we assign a *total* or *integral curvature*, which is represented by the area of the figure on the sphere corresponding to it. From this integral curvature must be distinguished the somewhat more specific curvature which we shall call the *measure of curvature*. The latter refers to a *point* of the surface, and shall denote the quotient obtained when the integral curvature of the surface element about a point is divided by the area of the element itself; and hence it denotes the ratio of the infinitely small areas which correspond to one another on the curved surface and on the sphere. The use of these innovations will be abundantly justified, as we hope, by what we shall explain below. As for the terminology, we have thought it especially desirable that all ambiguity be avoided. For this reason we have not thought it advantageous to follow strictly the analogy of the terminology commonly adopted (though not approved by all) in the theory of plane curves, according to which the measure of curvature should be called simply curvature, but the total curvature, the amplitude. But why not be free in the choice of words, provided they are not meaningless and not liable to a misleading interpretation?

The position of a figure on the sphere can be either similar to the position of the corresponding figure on the curved surface, or opposite (inverse). The former is the case when two lines going out on the curved surface from the same point in different, but not opposite directions, are represented on the sphere by lines similarly placed, that is, when the map of the line to the right is also to the right; the latter is the case when the contrary holds. We shall distinguish these two cases by the positive or negative *sign* of the measure of curvature. But evidently this distinction can hold only when on each surface we choose a definite face on which we suppose the figure to lie. On the auxiliary sphere we shall use always the exterior face, that is, that turned away from the center; on the curved surface also there may be taken for the exterior face the one already considered, or rather that face from which the normal is supposed to be drawn. For, evidently, there is no change in regard to the similitude of the figures, if on the curved surface both the figure and the normal be transferred to the opposite side, so long as the image itself is represented on the same side of the sphere.

The positive or negative sign, which we assign to the *measure* of curvature according to the position of the infinitely small figure, we extend also to the

§7. In this section Gauss finds a formula for the curvature k at A . His answer, at the top of page 77, is given for a surface which is the graph of $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$. In this case, the functions X and Y can be thought of as functions on \mathbb{R}^2 (that is, we consider $X \circ f$ and $Y \circ f$), and Gauss' answer is

$$k = \frac{\partial X}{\partial x} \frac{\partial Y}{\partial y} - \frac{\partial X}{\partial y} \frac{\partial Y}{\partial x} \quad [= D_1(X \circ f)D_2(Y \circ f) - D_2(X \circ f)D_1(Y \circ f)].$$

[The notation $\frac{dX}{dx}$, etc., in the original has been preserved. Similarly, a few

integral curvature of a finite figure on the curved surface. However, if we wish to discuss the general case, some explanations will be necessary, which we can only touch here briefly. So long as the figure on the curved surface is such that to *distinct* points on itself there correspond distinct points on the sphere, the definition needs no further explanation. But whenever this condition is not satisfied, it will be necessary to take into account twice or several times certain parts of the figure on the sphere. Whence for a similar, or inverse position, may arise an accumulation of areas, or the areas may partially or wholly destroy each other. In such a case, the simplest way is to suppose the curved surface divided into parts, such that each part, considered separately, satisfies the above condition; to assign to each of the parts its integral curvature, determining this magnitude by the area of the corresponding figure on the sphere, and the sign by the position of this figure; and, finally, to assign to the total figure the integral curvature arising from the addition of the integral curvatures which correspond to the single parts. So, generally, the integral curvature of a figure is equal to $\int k d\sigma$, $d\sigma$ denoting the element of area of the figure, and k the measure of curvature at any point. The principal points concerning the geometric representation of this integral reduce to the following. To the perimeter of the figure on the curved surface (under the restriction of Art. 3) will correspond always a closed line on the sphere. If the latter nowhere intersect itself, it will divide the whole surface of the sphere into two parts, one of which will correspond to the figure on the curved surface; and its area (taken as positive or negative according as, with respect to its perimeter, its position is similar, or inverse, to the position of the figure on the curved surface) will represent the integral curvature of the figure on the curved surface. But whenever this line intersects itself once or several times, it will give a complicated figure, to which, however, it is possible to assign a definite area as legitimately as in the case of a figure without nodes; and this area, properly interpreted, will give always an exact value for the integral curvature. However, we must reserve for another occasion the more extended exposition of the theory of these figures viewed from this very general standpoint.

7.

We shall now find a formula which will express the measure of curvature for any point of a curved surface. Let $d\sigma$ denote the area of an element of this surface; then $Z d\sigma$ will be the area of the projection of this element on the plane of the coordinates x, y ; and consequently, if $d\Sigma$ is the area of the corresponding element on the sphere, $Z d\Sigma$ will be the area of its projection on the same plane. The positive or negative sign of Z will, in fact, indicate that the position of the projection is similar or inverse to that of the projected element. Evidently these

lines later the expressions $\frac{ddz}{dx^2}$ and $\frac{ddz}{dx \cdot dy}$ stand for $\frac{\partial^2 z}{\partial x^2}$ and $\frac{\partial^2 z}{\partial x \partial y}$, etc.]

Gauss obtains this answer by considering an infinitesimal triangle “ $d\sigma$ ” with one vertex at $(x, y, f(x, y))$, one vertex at $(x + dx, y + dy, f(x + dx, y + dy))$, and one at $(x + \delta x, y + \delta y, f(x + \delta x, y + \delta y))$. It is a challenge both to follow Gauss’ reasoning, and to put it in modern terms. Either way, one needs Gauss’ preliminary observation that

$$\frac{\text{area } v(d\sigma)}{\text{area } d\sigma} = \frac{\text{area of projection on } (x, y)\text{-plane of } v(d\sigma)}{\text{area of projection on } (x, y)\text{-plane of } d\sigma}.$$

This mysterious equation really says that the tangent plane of M at A is parallel to the tangent plane of S^2 at $v(A)$. If this hint does not help, simply accept the formula for k , which will be derived later, using modern terminology.

The remainder of section 7 evaluates k in terms of partial derivatives of f (which Gauss denotes by z).

projections have the same ratio as to quantity and the same relation as to position as the elements themselves. Let us consider now a triangular element on the curved surface, and let us suppose that the coordinates of the three points which form its projection are

$$\begin{array}{ll} x, & y \\ x + dx, & y + dy \\ x + \delta x, & y + \delta y \end{array}$$

The double area of this triangle will be expressed by the formula

$$dx \cdot \delta y - dy \cdot \delta x$$

and this will be in a positive or negative form according as the position of the side from the first point to the third, with respect to the side from the first point to the second, is similar or opposite to the position of the y -axis of coordinates with respect to the x -axis of coordinates.

In like manner, if the coordinates of the three points which form the projection of the corresponding element on the sphere, from the center of the sphere as origin, are

$$\begin{array}{ll} X, & Y \\ X + dX, & Y + dY \\ X + \delta X, & Y + \delta Y \end{array}$$

the double area of this projection will be expressed by

$$dX \cdot \delta Y - dY \cdot \delta X$$

and the sign of this expression is determined in the same manner as above. Wherefore the measure of curvature at this point of the curved surface will be

$$k = \frac{dX \cdot \delta Y - dY \cdot \delta X}{dx \cdot \delta y - dy \cdot \delta x}$$

If now we suppose the nature of the curved surface to be defined according to the third method considered in Art. 4, X and Y will be in the form of functions of the quantities x, y . We shall have, therefore,

$$\begin{aligned} dX &= \left(\frac{dX}{dx}\right) dx + \left(\frac{dX}{dy}\right) dy \\ \delta X &= \left(\frac{dX}{dx}\right) \delta x + \left(\frac{dX}{dy}\right) \delta y \\ dY &= \left(\frac{dY}{dx}\right) dx + \left(\frac{dY}{dy}\right) dy \\ \delta Y &= \left(\frac{dY}{dx}\right) \delta x + \left(\frac{dY}{dy}\right) \delta y \end{aligned}$$

When these values have been substituted, the above expression becomes

$$k = \left(\frac{dX}{dx}\right)\left(\frac{dY}{dy}\right) - \left(\frac{dX}{dy}\right)\left(\frac{dY}{dx}\right)$$

Setting, as above,

$$\frac{dz}{dx} = t, \quad \frac{dz}{dy} = u$$

and also

$$\frac{ddz}{dx^2} = T, \quad \frac{ddz}{dx \cdot dy} = U, \quad \frac{ddz}{dy^2} = V$$

or

$$dt = T dx + U dy, \quad du = U dx + V dy$$

we have from the formulæ given above

$$X = -tZ, \quad Y = -uZ, \quad (1 + t^2 + u^2)Z^2 = 1$$

and hence

$$dX = -Z dt - t dZ$$

$$dY = -Z du - u dZ$$

$$(1 + t^2 + u^2) dZ + Z(t dt + u du) = 0$$

or

$$dZ = -Z^3(t dt + u du)$$

$$dX = -Z^3(1 + u^2) dt + Z^3 t u du$$

$$dY = +Z^3 t u dt - Z^3(1 + t^2) du$$

and so

$$\frac{dX}{dx} = Z^3(- (1 + u^2)T + t u U)$$

$$\frac{dX}{dy} = Z^3(- (1 + u^2)U + t u V)$$

$$\frac{dY}{dx} = Z^3(t u T - (1 + t^2)U)$$

$$\frac{dY}{dy} = Z^3(t u U - (1 + t^2)V)$$

Substituting these values in the above expression, it becomes

$$k = Z^6(TV - U^2)(1 + t^2 + u^2) = Z^4(TV - U^2) = \frac{TV - U^2}{(1 + t^2 + u^2)^2}$$

§8. This section, except for the last theorem, was already done in Chapter 2.

8.

By a suitable choice of origin and axes of coordinates, we can easily make the values of the quantities t , u , U vanish for a definite point A . Indeed, the first two conditions will be fulfilled at once if the tangent plane at this point be taken for the xy -plane. If, further, the origin is placed at the point A itself, the expression for the coordinate z evidently takes the form

$$z = \frac{1}{2}T^0x^2 + U^0xy + \frac{1}{2}V^0y^2 + \Omega$$

where Ω will be of higher degree than the second. Turning now the axes of x and y through an angle M such that

$$\tan 2M = \frac{2U^0}{T^0 - V^0}$$

it is easily seen that there must result an equation of the form

$$z = \frac{1}{2}Tx^2 + \frac{1}{2}Vy^2 + \Omega$$

In this way the third condition is also satisfied. When this has been done, it is evident that

I. If the curved surface be cut by a plane passing through the normal itself and through the x -axis, a plane curve will be obtained, the radius of curvature of which at the point A will be $= \frac{1}{T}$, the positive or negative sign indicating that the curve is concave or convex toward that region toward which the coordinates z are positive.

II. In like manner $\frac{1}{V}$ will be the radius of curvature at the point A of the plane curve which is the intersection of the surface and the plane through the y -axis and the z -axis.

III. Setting $x = r \cos \varphi$, $y = r \sin \varphi$, the equation becomes

$$z = \frac{1}{2}(T \cos^2 \varphi + V \sin^2 \varphi)r^2 + \Omega$$

from which we see that if the section is made by a plane through the normal at A and making an angle φ with the x -axis, we shall have a plane curve whose radius of curvature at the point A will

$$= \frac{1}{T \cos^2 \varphi + V \sin^2 \varphi}$$

IV. Therefore, whenever we have $T = V$, the radii of curvature in *all* the normal planes will be equal. But if T and V are not equal, it is evident that, since

§§9, 10, 11. These sections are essentially calculations, involving no new ideas. Every once in a while Gauss calculates a differential instead of some partial derivatives, but this should cause no difficulties.

The goal is the very last, four-line-long, equation at the end of section 11.

for any value whatever of the angle φ , $T \cos^2 \varphi + V \sin^2 \varphi$ falls between T and V , the radii of curvature in the principal sections considered in I and II refer to the extreme curvatures; that is to say, the one to the maximum curvature, the other to the minimum, if T and V have the same sign. On the other hand, one has the greatest convex curvature, the other the greatest concave curvature, if T and V have opposite signs. These conclusions contain almost all that the illustrious EULER was the first to prove on the curvature of curved surfaces.

V. The measure of curvature at the point A on the curved surface takes the very simple form $k = TV$, whence we have the

THEOREM. *The measure of curvature at any point whatever of the surface is equal to a fraction whose numerator is unity, and whose denominator is the product of the two extreme radii of curvature of the sections by normal planes.*

At the same time it is clear that the measure of curvature is positive for concavo-concave or convexo-convex surfaces (which distinction is not essential), but negative for concavo-convex surfaces. If the surface consists of parts of each kind, then on the lines separating the two kinds the measure of curvature ought to vanish. Later we shall make a detailed study of the nature of curved surfaces for which the measure of curvature everywhere vanishes.

9.

The general formula for the measure of curvature given at the end of Art. 7 is the most simple of all, since it involves only five elements. We shall arrive at a more complicated formula, indeed, one involving nine elements, if we wish to use the first method of representing a curved surface. Keeping the notation of Art. 4, let us set also

$$\begin{aligned} \frac{ddW}{dx^2} &= P', & \frac{ddW}{dy^2} &= Q', & \frac{ddW}{dx^2} &= R' \\ \frac{ddW}{dy \cdot dz} &= P'', & \frac{ddW}{dx \cdot dz} &= Q'', & \frac{ddW}{dx \cdot dy} &= R'' \end{aligned}$$

so that

$$\begin{aligned} dP &= P' dx + R'' dy + Q'' dz \\ dQ &= R'' dx + Q' dy + P'' dz \\ dR &= Q'' dx + P'' dy + R' dz \end{aligned}$$

Now since $t = -\frac{P}{R}$, we find through differentiation

$$R^2 dt = -R dP + P dR = (PQ'' - RP') dx + (PP'' - RR'') dy + (PR' - RQ'') dz$$

[As you can probably figure out for yourself, \mathfrak{c} is an alternate form of β .]

or, eliminating dz by means of the equation $P dx + Q dy + R dz = 0$,

$$R^3 dt = (-R^2 P' + 2PRQ'' - P^2 R') dx + (PRP'' + QRQ'' - PQR' - R^2 R'') dy$$

In like manner we obtain

$$R^3 du = (PRP'' + QRQ'' - PQR' - R^2 R'') dx + (-R^2 Q' + 2QRP'' - Q^2 R') dy$$

From this we conclude that

$$\begin{aligned} R^3 T &= -R^2 P' + 2PRQ'' - P^2 R' \\ R^3 U &= PRP'' + QRQ'' - PQR' - R^2 R'' \\ R^3 V &= -R^2 Q' + 2QRP'' - Q^2 R' \end{aligned}$$

Substituting these values in the formula of Art. 7, we obtain for the measure of curvature k the following symmetric expression:

$$\begin{aligned} (P^2 + Q^2 + R^2)^2 k &= P^2(Q'R' - P''^2) + Q^2(P'R' - Q''^2) + R^2(P'Q' - R''^2) \\ &\quad + 2QR(Q''R'' - P'P'') + 2PR(P''R'' - Q'Q'') + 2PQ(P''Q'' - R'R'') \end{aligned}$$

10.

We obtain a still more complicated formula, indeed, one involving fifteen elements, if we follow the second general method of defining the nature of a curved surface. It is, however, very important that we develop this formula also. Retaining the notations of Art. 4, let us put also

$$\begin{aligned} \frac{ddx}{dp^2} &= \alpha, & \frac{ddx}{dp \cdot dq} &= \alpha', & \frac{ddx}{dq^2} &= \alpha'' \\ \frac{ddy}{dp^2} &= \beta, & \frac{ddy}{dp \cdot dq} &= \beta', & \frac{ddy}{dq^2} &= \beta'' \\ \frac{ddz}{dp^2} &= \gamma, & \frac{ddz}{dp \cdot dq} &= \gamma', & \frac{ddz}{dq^2} &= \gamma'' \end{aligned}$$

and let us put, for brevity,

$$\begin{aligned} bc' - cb' &= A \\ ca' - ac' &= B \\ ab' - ba' &= C \end{aligned}$$

First we see that $A \, dx + B \, dy + C \, dz = 0$, or $dz = -\frac{A}{C} \, dx - \frac{B}{C} \, dy$; thus, inasmuch as z may be regarded as a function of x, y , we have

$$\begin{aligned}\frac{dz}{dx} &= t = -\frac{A}{C} \\ \frac{dz}{dy} &= u = -\frac{B}{C}\end{aligned}$$

Then from the formulæ $dx = a \, dp + a' \, dq$, $dy = b \, dp + b' \, dq$, we have

$$\begin{aligned}C \, dp &= b' \, dx - a' \, dy \\ C \, dq &= -b \, dx + a \, dy\end{aligned}$$

Thence we obtain for the total differentials of t, u

$$\begin{aligned}C^3 \, dt &= (A \frac{dC}{dp} - C \frac{dA}{dp})(b' \, dx - a' \, dy) + (C \frac{dA}{dq} - A \frac{dC}{dq})(b \, dx - a \, dy) \\ C^3 \, du &= (B \frac{dC}{dp} - C \frac{dB}{dp})(b' \, dx - a' \, dy) + (C \frac{dB}{dq} - B \frac{dC}{dq})(b \, dx - a \, dy)\end{aligned}$$

If now we substitute in these formulæ

$$\begin{aligned}\frac{dA}{dp} &= c' \epsilon + b \gamma' - c \epsilon' - b' \gamma \\ \frac{dA}{dq} &= c' \epsilon' + b \gamma'' - c \epsilon'' - b' \gamma' \\ \frac{dB}{dp} &= a' \gamma + c \alpha' - a \gamma' - c' \alpha \\ \frac{dB}{dq} &= a' \gamma' + c \alpha'' - a \gamma'' - c' \alpha' \\ \frac{dC}{dp} &= b' \alpha + a \epsilon' - b \alpha' - a' \epsilon \\ \frac{dC}{dq} &= b' \alpha' + a \epsilon'' - b \alpha'' - a' \epsilon'\end{aligned}$$

and if we note that the values of the differentials dt, du thus obtained must be equal, independently of the differentials dx, dy , to the quantities $T \, dx + U \, dy$, $U \, dx + V \, dy$ respectively, we shall find, after some sufficiently obvious

transformations,

$$\begin{aligned}
 C^3T &= \alpha Ab'^2 + 6Bb'^2 + \gamma Cb'^2 \\
 &\quad - 2\alpha' Abb' - 26'Bbb' - 2\gamma' Cbb' \\
 &\quad + \alpha'' Ab^2 + 6''Bb^2 + \gamma'' Cb^2 \\
 C^3U &= -\alpha Aa'b' - 6Ba'b' - \gamma Ca'b' \\
 &\quad + \alpha' A(ab' + ba') + 6'B(ab' + ba') + \gamma' C(ab' + ba') \\
 &\quad - \alpha'' Aab - 6''Bab - \gamma'' Cab \\
 C^3V &= \alpha Aa'^2 + 6Ba'^2 + \gamma Ca'^2 \\
 &\quad - 2\alpha' Aaa' - 26'Baa' - 2\gamma' Caa' \\
 &\quad + \alpha'' Aa^2 + 6''Ba^2 + \gamma'' Ca^2
 \end{aligned}$$

Hence, if we put, for the sake of brevity,

$$A\alpha + B6 + C\gamma = D \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

$$A\alpha' + B6' + C\gamma' = D' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$A\alpha'' + B6'' + C\gamma'' = D'' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

we shall have

$$\begin{aligned}
 C^3T &= Db'^2 - 2D'bb' + D''b^2 \\
 C^3U &= -Da'b' + D'(ab' + ba') - D''ab \\
 C^3V &= Da'^2 - 2D'aa' + D''a^2
 \end{aligned}$$

From this we find, after the reckoning has been carried out,

$$C^6(TV - U^2) = (DD'' - D'^2)(ab' - ba')^2 = (DD'' - D'^2)C^2$$

and therefore the formula for the measure of curvature

$$k = \frac{DD'' - D'^2}{(A^2 + B^2 + C^2)^2}$$

By means of the formula just found we are going to establish another, which may be counted among the most productive theorems in the theory of curved

surfaces. Let us introduce the following notation:

$$\begin{aligned} a^2 + b^2 + c^2 &= E \\ aa' + bb' + cc' &= F \\ a'^2 + b'^2 + c'^2 &= G \\ a\alpha + b\beta + c\gamma &= m \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4) \end{aligned}$$

$$a\alpha' + b\beta' + c\gamma' = m' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

$$a\alpha'' + b\beta'' + c\gamma'' = m'' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$a'\alpha + b'\epsilon + c'\gamma = n \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

$$a'\alpha' + b'\beta' + c'\gamma' = n' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

$$a'\alpha'' + b'\beta'' + c'\gamma'' = n'' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$A^2 + B^2 + C^2 = EG - F^2 = \Delta$$

Let us eliminate from the equations 1, 4, 7 the quantities δ, γ , which is done by multiplying them by $bc' - cb'$, $b'C - c'B$, $cB - bC$ respectively and adding: in this way we obtain

$$\begin{aligned} & (A(bc' - cb') + a(b'C - c'B) + a'(cB - bC))\alpha \\ & = D(bc' - cb') + m(b'C - c'B) + n(cB - bC) \end{aligned}$$

an equation which is easily transformed into

$$AD = \alpha\Delta + a(nF - mG) + a'(mF - nE)$$

Likewise the elimination of α , γ or α , δ from the same equations gives

$$BD = 6\Delta + b(nF - mG) + b'(mF - nE)$$

$$CD = \gamma\Delta + c(nF - mG) + c'(mF - nE)$$

Multiplying these three equations by α'' , β'' , γ'' respectively and adding, we obtain

$$DD'' = (\alpha\alpha'' + \ell\ell'' + \gamma\gamma'')\Delta + m''(nF - mG) + n''(mF - nE) . \quad (10)$$

If we treat the equations 2, 5, 8 in the same way, we obtain

$$AD' = \alpha' \Delta + a(n'F - m'G) + a'(m'F - n'E)$$

$$BD' = \epsilon' \Delta + b(n'F - m'G) + b'(m'F - n'E)$$

$$CD' = \gamma' \Delta + c(n'F - m'G) + c'(m'F - n'E)$$

§12. If $M, N \subset \mathbb{R}^3$ are surfaces, then a *development* of M on N is simply a map $f: M \rightarrow N$ which is an isometry (with respect to the induced Riemannian metrics).

and after these equations are multiplied by α' , β' , γ' respectively, addition gives

$$D'^2 = (\alpha'^2 + \beta'^2 + \gamma'^2)\Delta + m'(n'F - m'G) + n'(m'F - n'E)$$

A combination of this equation with equation (10) gives

$$\begin{aligned} DD'' - D'^2 &= (\alpha\alpha'' + \beta\beta'' + \gamma\gamma'' - \alpha'^2 - \beta'^2 - \gamma'^2)\Delta \\ &\quad + E(n'^2 - nn'') + F(nm'' - 2m'n' + mn'') + G(m'^2 - mm'') \end{aligned}$$

It is clear that we have

$$\frac{dE}{dp} = 2m, \quad \frac{dE}{dq} = 2m', \quad \frac{dF}{dp} = m' + n, \quad \frac{dF}{dq} = m'' + n', \quad \frac{dG}{dp} = 2n', \quad \frac{dG}{dq} = 2n''$$

or

$$\begin{aligned} m &= \frac{1}{2} \frac{dE}{dp}, & m' &= \frac{1}{2} \frac{dE}{dq}, & m'' &= \frac{dF}{dq} - \frac{1}{2} \frac{dG}{dp} \\ n &= \frac{dF}{dp} - \frac{1}{2} \frac{dE}{dq}, & n' &= \frac{1}{2} \frac{dG}{dp}, & n'' &= \frac{1}{2} \frac{dG}{dq} \end{aligned}$$

Moreover, it is easily shown that we shall have

$$\begin{aligned} \alpha\alpha'' + \beta\beta'' + \gamma\gamma'' - \alpha'^2 - \beta'^2 - \gamma'^2 &= \frac{dn}{dq} - \frac{dn'}{dp} = \frac{dm''}{dp} - \frac{dm'}{dq} \\ &= -\frac{1}{2} \cdot \frac{ddE}{dq^2} + \frac{ddF}{dp \cdot dq} - \frac{1}{2} \cdot \frac{ddG}{dp^2} \end{aligned}$$

If we substitute these different expressions in the formula for the measure of curvature derived at the end of the preceding article, we obtain the following formula, which involves only the quantities E , F , G and their differential quotients of the first and second orders:

$$\begin{aligned} 4(EG - F^2)^2 k &= E \left(\frac{dE}{dq} \cdot \frac{dG}{dq} - 2 \frac{dF}{dp} \cdot \frac{dG}{dq} + \left(\frac{dG}{dp} \right)^2 \right) \\ &\quad + F \left(\frac{dE}{dp} \cdot \frac{dG}{dq} - \frac{dE}{dq} \cdot \frac{dG}{dp} - 2 \frac{dE}{dq} \cdot \frac{dF}{dq} + 4 \frac{dF}{dp} \cdot \frac{dF}{dq} - 2 \frac{dF}{dp} \cdot \frac{dG}{dp} \right) \\ &\quad + G \left(\frac{dE}{dp} \cdot \frac{dG}{dp} - 2 \frac{dE}{dp} \cdot \frac{dF}{dq} + \left(\frac{dE}{dq} \right)^2 \right) \\ &\quad - 2(EG - F^2) \left(\frac{ddE}{dq^2} - 2 \frac{ddF}{dp \cdot dq} + \frac{ddG}{dp^2} \right) \end{aligned}$$

12.

Since we always have

$$dx^2 + dy^2 + dz^2 = E dp^2 + 2F dp \cdot dq + G dq^2$$

it is clear that $\sqrt{(E dp^2 + 2F dp \cdot dq + G dq^2)}$ is the general expression for the linear element on the curved surface. The analysis developed in the preceding article thus shows us that for finding the measure of curvature there is no need of finite formulæ, which express the coordinates x, y, z as functions of the indeterminants p, q ; but that the general expression for the magnitude of any linear element is sufficient. Let us proceed to some applications of this very important theorem.

Suppose that our surface can be developed upon another surface, curved or plane, so that to each point of the former surface, determined by the coordinates x, y, z , will correspond a definite point of the latter surface, whose coordinates are x', y', z' . Evidently x', y', z' can also be regarded as functions of the indeterminants p, q , and therefore for the element $\sqrt{(dx'^2 + dy'^2 + dz'^2)}$ we shall have an expression of the form

$$\sqrt{(E' dp^2 + 2F' dp \cdot dq + G' dq^2)}$$

where E', F', G' also denote functions of p, q . But from the very notion of the *development* of one surface upon another it is clear that the elements corresponding to one another on the two surfaces are necessarily equal. Therefore we shall have identically

$$E = E', \quad F = F', \quad G = G'$$

Thus the formula of the preceding article leads of itself to the remarkable

THEOREM. *If a curved surface is developed upon any other surface whatever, the measure of curvature in each point remains unchanged.*

Also it is evident that *any finite part whatever of the curved surface will retain the same integral curvature after development upon another surface.*

Surfaces developable upon a plane constitute the particular case to which geometers have heretofore restricted their attention. Our theory shows at once that the measure of curvature at every point of such surfaces is equal to zero. Consequently, if the nature of these surfaces is defined according to the third method, we shall have at every point

$$\frac{ddz}{dx^2} \cdot \frac{ddz}{dy^2} - \left(\frac{ddz}{dx \cdot dy} \right)^2 = 0$$

a criterion which, though indeed known a short time ago, has not, at least to our knowledge, commonly been demonstrated with as much rigor as is desirable.

What we have explained in the preceding article is connected with a particular method of studying surfaces, a very worthy method which may be thoroughly

§14. Throughout this section Gauss uses x, y, z to denote $x \circ c, y \circ c, z \circ c$, for the curve c under consideration. The integral in the second display, involving both d and δ , is what we would write as

$$\begin{aligned} \left. \frac{dL(\bar{\alpha}(u))}{du} \right|_{u=0} &= \left. \frac{d}{du} \right|_{u=0} \int_a^b \sqrt{\left(\frac{\partial \alpha^1(u, t)}{\partial t} \right)^2 + \dots} dt \\ &= \int_a^b \frac{\frac{\partial \alpha^1(0, t)}{\partial u} \frac{\partial^2 \alpha^1(0, t)}{\partial u \partial t} + \dots}{\sqrt{\left(\frac{\partial \alpha^1(0, t)}{\partial u} \right)^2 + \dots}} dt = \int_a^b \frac{\frac{dc^1}{dt} \frac{\partial^2 \alpha^1(0, t)}{\partial u \partial t} + \dots}{\sqrt{\left(\frac{\partial \alpha^1(0, t)}{\partial u} \right)^2 + \dots}} dt. \end{aligned}$$

Thus, $\frac{dc^1}{dt}$ is $\frac{dx}{[dt]}$ and $\frac{\partial^2 \alpha^1(0, t)}{\partial u \partial t} = \frac{\partial^2 \alpha^1(0, t)}{\partial t \partial u}$ is $\frac{d\delta x}{[\partial t \partial u]}$. The next two lines show what this becomes after integration by parts. The integral is

$$- \int_a^b \frac{\partial \alpha^1}{\partial u}(0, t) \frac{d}{dt} \left(\frac{dc^1/dt}{\sqrt{\dots}} \right) + \dots dt;$$

here $\frac{\partial \alpha^1}{\partial u}(0, t)$ is $\frac{\delta x}{[\partial u]}$.

developed by geometers. When a surface is regarded, not as the boundary of a solid, but as a flexible, though not extensible solid, one dimension of which is supposed to vanish, then the properties of the surface depend in part upon the form to which we can suppose it reduced, and in part are absolute and remain invariable, whatever may be the form into which the surface is bent. To these latter properties, the study of which opens to geometry a new and fertile field, belong the measure of curvature and the integral curvature, in the sense which we have given to these expressions. To these belong also the theory of shortest lines, and a great part of what we reserve to be treated later. From this point of view, a plane surface and a surface developable on a plane, e.g., cylindrical surfaces, conical surfaces, etc., are to be regarded as essentially identical; and the generic method of defining in a general manner the nature of the surfaces thus considered is always based upon the formula $\sqrt{(E dp^2 + 2F dp \cdot dq + G dq^2)}$, which connects the linear element with the two indeterminants p, q . But before following this study further, we must introduce the principles of the theory of shortest lines on a given curved surface.

14.

The nature of a curved line in space is generally given in such a way that the coordinates x, y, z corresponding to the different points of it are given in the form of functions of a single variable, which we shall call w . The length of such a line from an arbitrary initial point to the point whose coordinates are x, y, z , is expressed by the integral

$$\int dw \cdot \sqrt{\left(\frac{dx}{dw}\right)^2 + \left(\frac{dy}{dw}\right)^2 + \left(\frac{dz}{dw}\right)^2}$$

If we suppose that the position of the line undergoes an infinitely small variation, so that the coordinates of the different points receive the variations $\delta x, \delta y, \delta z$, the variation of the whole length becomes

$$= \int \frac{dx \cdot d\delta x + dy \cdot d\delta y + dz \cdot d\delta z}{\sqrt{(dx^2 + dy^2 + dz^2)}}$$

which expression we can change into the form

$$\frac{dx \cdot \delta x + dy \cdot \delta y + dz \cdot \delta z}{\sqrt{(dx^2 + dy^2 + dz^2)}} - \int \left(\delta x \cdot d \frac{dx}{\sqrt{(dx^2 + dy^2 + dz^2)}} + \delta y \cdot d \frac{dy}{\sqrt{(dx^2 + dy^2 + dz^2)}} + \delta z \cdot d \frac{dz}{\sqrt{(dx^2 + dy^2 + dz^2)}} \right)$$

We know that, in case the line is to be the shortest between its end points, all that stands under the integral sign must vanish. Since the line must lie on the given

Notice that Gauss has given $dL(\bar{\alpha}(u))/du|_{u=0}$ for an arbitrary variation in 3-space, not just a variation through curves in the surface. His $x = x \circ c$ is a coordinate function of c in 3-space, *not* a coordinate function with respect to some coordinate system on the surface. If the surface is $\{p : W(p) = 0\}$ for some $W: \mathbb{R}^3 \rightarrow \mathbb{R}$, so that on the surface we have

$$0 = dW = P dx + Q dy + R dz \quad P = D_1 W, \quad Q = D_2 W, \quad R = D_3 W,$$

then for variations α through curves on the surface we will have

$$dW(\delta x, \delta y, \delta z) = dW \left(\frac{\partial \alpha^1}{\partial u}(0, t), \frac{\partial \alpha^2}{\partial u}(0, t), \frac{\partial \alpha^3}{\partial u}(0, t) \right) = 0,$$

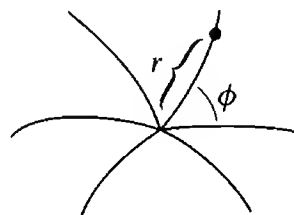
and any set of $\partial \alpha^i / \partial u(0, t)$ with this property comes from some variation on the surface. Using this, Gauss deduces a necessary and sufficient condition for a curve γ , parameterized by arclength, to be a geodesic on the surface. Unlike our equations for geodesics, this condition [the next-to-last displayed formula in this section] is expressed in terms of quantities which make sense only in \mathbb{R}^3 :

$$\frac{\gamma^{1''}(t)}{X(\gamma(t))} = \frac{\gamma^{2''}(t)}{Y(\gamma(t))} = \frac{\gamma^{3''}(t)}{Z(\gamma(t))},$$

i.e., $\gamma''(t)$ is a multiple of the normal vector at $\gamma(t)$. It takes a little detective work to see that Gauss is really considering a curve parameterized by arclength. Try to prove Gauss' result by modifying our proof of Euler's equations.

§15. The proof in this section is essentially our (first) proof of Gauss' Lemma (I.9-12). There are two main differences. First, Gauss uses the condition of section 14 rather than our equations. Second, for a surface it is unnecessary to choose a curve $v: \mathbb{R} \rightarrow M_q$ and manufacture the variation α that occurs in the proof of Lemma I.9-12. Instead, we just use

$\alpha(r, \phi) = \text{point with "polar coordinates" } (r, \phi).$



surface, whose nature is defined by the equation $P dx + Q dy + R dz = 0$, the variations $\delta x, \delta y, \delta z$ also must satisfy the equation $P \delta x + Q \delta y + R \delta z = 0$, and from this it follows at once, according to well-known rules, that the differentials

$$d \frac{dx}{\sqrt{(dx^2 + dy^2 + dz^2)}}, \quad d \frac{dy}{\sqrt{(dx^2 + dy^2 + dz^2)}}, \quad d \frac{dz}{\sqrt{(dx^2 + dy^2 + dz^2)}}$$

must be proportional to the quantities P, Q, R respectively. Let dr be the element of the curved line; λ the point on the sphere representing the direction of this element; L the point on the sphere representing the direction of the normal to the curved surface; finally, let ξ, η, ζ be the coordinates of the point λ , and X, Y, Z be those of the point L with reference to the center of the sphere. We shall then have

$$dx = \xi dr, \quad dy = \eta dr, \quad dz = \zeta dr$$

from which we see that the above differentials become $d\xi, d\eta, d\zeta$. And since the quantities P, Q, R are proportional to X, Y, Z , the character of shortest lines is expressed by the equations

$$\frac{d\xi}{X} = \frac{d\eta}{Y} = \frac{d\zeta}{Z}$$

Moreover, it is easily seen that $\sqrt{(d\xi^2 + d\eta^2 + d\zeta^2)}$ is equal to the small arc on the sphere which measures the angle between the directions of the tangents at the beginning and at the end of the element dr , and is thus $= \frac{dr}{\rho}$, if ρ denotes the radius of curvature of the shortest line at this point; thus we shall have

$$\rho d\xi = X dr, \quad \rho d\eta = Y dr, \quad \rho d\zeta = Z dr$$

15.

Suppose that an infinite number of shortest lines go out from a given point A on the curved surface, and suppose that we distinguish these lines from one another by the angle that the first element of each of them makes with the first element of one of them which we take for the first. Let φ be that angle, or, more generally, a function of that angle, and r the length of such a shortest line from the point A to the point whose coordinates are x, y, z . Since to definite values of the variables r, φ there correspond definite points of the surface, the coordinates x, y, z can be regarded as function of r, φ . We shall retain for the notation $\lambda, L, \xi, \eta, \zeta, X, Y, Z$ the same meaning as in the preceding article, this notation referring to any point whatever on any one of the shortest lines.

All the shortest lines that are of the same length r will end on another line whose length, measured from an arbitrary initial point, we shall denote by v . Thus v can be regarded as a function of the indeterminants r, φ , and if λ' denotes the point on the sphere corresponding to the direction of the element dv , and also ξ', η', ζ' denote the coordinates of this point with reference to the center of the sphere, we shall have

$$\frac{dx}{d\varphi} = \xi' \cdot \frac{dv}{d\varphi}, \quad \frac{dy}{d\varphi} = \eta' \cdot \frac{dv}{d\varphi}, \quad \frac{dz}{d\varphi} = \zeta' \cdot \frac{dv}{d\varphi}$$

From these equations and from the equations

$$\frac{dx}{dr} = \xi, \quad \frac{dy}{dr} = \eta, \quad \frac{dz}{dr} = \zeta$$

we have

$$\frac{dx}{dr} \cdot \frac{dx}{d\varphi} + \frac{dy}{dr} \cdot \frac{dy}{d\varphi} + \frac{dz}{dr} \cdot \frac{dz}{d\varphi} = (\xi\xi' + \eta\eta' + \zeta\zeta') \cdot \frac{dv}{d\varphi} = \cos \lambda\lambda' \cdot \frac{dv}{d\varphi}$$

Let S denote the first member of this equation, which will also be a function of r, φ . Differentiation of S with respect to r gives

$$\begin{aligned} \frac{dS}{dr} &= \frac{d^2x}{dr^2} \cdot \frac{dx}{d\varphi} + \frac{d^2y}{dr^2} \cdot \frac{dy}{d\varphi} + \frac{d^2z}{dr^2} \cdot \frac{dz}{d\varphi} + \frac{1}{2} \cdot \frac{d\left(\left(\frac{dx}{dr}\right)^2 + \left(\frac{dy}{dr}\right)^2 + \left(\frac{dz}{dr}\right)^2\right)}{d\varphi} \\ &= \frac{d\xi}{dr} \cdot \frac{dx}{d\varphi} + \frac{d\eta}{dr} \cdot \frac{dy}{d\varphi} + \frac{d\zeta}{dr} \cdot \frac{dz}{d\varphi} + \frac{1}{2} \cdot \frac{d(\xi^2 + \eta^2 + \zeta^2)}{d\varphi} \end{aligned}$$

But $\xi^2 + \eta^2 + \zeta^2 = 1$, and therefore its differential = 0; and by the preceding article we have, if ρ denotes the radius of curvature of the line r ,

$$\frac{d\xi}{dr} = \frac{X}{\rho}, \quad \frac{d\eta}{dr} = \frac{Y}{\rho}, \quad \frac{d\zeta}{dr} = \frac{Z}{\rho}$$

Thus we have

$$\frac{dS}{dr} = \frac{1}{\rho} \cdot (X\xi' + Y\eta' + Z\zeta') \cdot \frac{dv}{d\varphi} = \frac{1}{\rho} \cdot \cos L\lambda' \cdot \frac{dv}{d\varphi} = 0$$

since λ' evidently lies on the great circle whose pole is L . From this we see that S is independent of r , and is, therefore, a function of φ alone. But for $r = 0$ we evidently have $v = 0$, consequently $\frac{dv}{d\varphi} = 0$, and $S = 0$ independently of φ . Thus, in general, we have necessarily $S = 0$, and so $\cos \lambda\lambda' = 0$, i.e., $\lambda\lambda' = 90^\circ$. From this follows the

Gauss also gives a “geometric” proof of the lemma, using infinitesimal triangles. Perhaps the easiest way to make this rigorous would be to use our second proof of Gauss’ Lemma.

§16. This section states a generalization of Gauss’ Lemma, which has also been given in Problem I.9-28.

§17. In terms of a coordinate system (p, q) on a surface, the Riemannian metric that it acquires as a subset of \mathbb{R}^3 has the expression

$$\langle \cdot, \cdot \rangle = E dp \otimes dp + F dp \otimes dq + F dq \otimes dp + G dq \otimes dq,$$

so that

$$\| \cdot \| = \sqrt{E dp \cdot dp + 2F dp \cdot dq + G dq \cdot dq}.$$

THEOREM. *If on a curved surface an infinite number of shortest lines of equal length be drawn from the same initial point, the lines joining their extremities will be normal to each of the lines.*

We have thought it worth while to deduce this theorem from the fundamental property of shortest lines; but the truth of the theorem can be made apparent without any calculation by means of the following reasoning. Let AB , AB' be two shortest lines of the same length including at A an infinitely small angle, and let us suppose that one of the angles made by the element BB' with the lines BA , $B'A$ differs from a right angle by a finite quantity. Then, by the law of continuity, one will be greater and the other less than a right angle. Suppose the angle at B is equal to $90^\circ - \omega$, and take on the line AB a point C , such that $BC = BB' \cdot \operatorname{cosec} \omega$. Then, since the infinitely small triangle $BB'C$ may be regarded as plane, we shall have $CB' = BC \cdot \cos \omega$, and consequently

$$AC + CB' = AC + BC \cdot \cos \omega = AB - BC \cdot (1 - \cos \omega) = AB' - BC \cdot (1 - \cos \omega),$$

i.e., the path from A to B' through the point C is shorter than the shortest line, Q.E.D.

16.

With the theorem of the preceding article we associate another, which we state as follows: *If on a curved surface we imagine any line whatever, from the different points of which are drawn at right angles and toward the same side an infinite number of shortest lines of the same length, the curve which joins their other extremities will cut each of the lines at right angles.* For the demonstration of this theorem no change need be made in the preceding analysis, except that φ must denote the length of the *given* curve measured from an arbitrary point; or rather, a function of this length. Thus all of the reasoning will hold here also, with this modification, that $S = 0$ for $r = 0$ is now implied in the hypothesis itself. Moreover, this theorem is more general than the preceding one, for we can regard it as including the first one if we take for the given line the infinitely small circle described about the center A . Finally, we may say that here also geometric considerations may take the place of the analysis, which, however, we shall not take the time to consider here, since they are sufficiently obvious.

17.

We return to the formula $\sqrt{(E dp^2 + 2F dp \cdot dq + G dq^2)}$, which expresses generally the magnitude of a linear element on the curved surface, and investigate, first of all, the geometric meaning of the coefficients E , F , G . We have

Gauss uses ω to denote the angle between $\partial/\partial p$ and $\partial/\partial q$ (thus, ω is a function on the surface). Gauss' formula for $\cos \omega$ should be clear. Gauss also mentions that

$$dV = \sqrt{EG - F^2} dp \wedge dq,$$

a special case of the formula on pg. I.311.

To interpret the last two formulas in this section, we must divide ds , dp , and dq by dt in all places; it is to be understood that $dp/dt = (p \circ c)'(t)$, etc., where c is the curve we are considering. It is simplest to assume that c is parameterized by arclength, so that the terms ds/dt are 1. If

$$\theta(s) = \text{angle between } c'(s) \text{ and } \left. \frac{\partial}{\partial p} \right|_{c(s)},$$

then

$$\cos \theta = \frac{\left\langle c', \frac{\partial}{\partial p} \right\rangle}{\left\| \frac{\partial}{\partial p} \right\|} = \frac{E \frac{dp(c(s))}{ds} + F \frac{dq(c(s))}{ds}}{\sqrt{E}},$$

since

$$c' = \frac{dp(c(s))}{ds} \frac{\partial}{\partial p} + \frac{dq(c(s))}{ds} \frac{\partial}{\partial q}.$$

Moreover, the area of the parallelogram spanned by c' and $\partial/\partial p$ is

$$\sin \theta \cdot \left\| \frac{\partial}{\partial p} \right\|, \quad \text{and also} \quad dV \left(\frac{\partial}{\partial p}, c' \right),$$

from which we obtain

$$\sin \theta = \frac{\sqrt{EG - F^2} \frac{dq(c(s))}{ds}}{\sqrt{E}}.$$

§18. In this section Gauss deduces the conditions for a curve γ (having the component functions $\gamma^1 = p \circ \gamma$, $\gamma^2 = q \circ \gamma$) to be a critical point for the length function.

Unlike the condition in section 14, the result is expressed totally in terms of the Riemannian metric $\langle \cdot, \cdot \rangle$ on the surface, and is essentially the condition for a geodesic that we obtained in Chapter I.9. However, the derivation is different, because the geodesic is assumed to satisfy $q(\gamma(t)) = \gamma^2(t) = t$ ["we regard p as a function of q "].

It is not necessary to actually follow the derivation. The really important point is simply the equation that constitutes the first line that appears in the

already said in Art. 5 that two systems of lines may be supposed to lie on the curved surface, p being variable, q constant along each of the lines of the one system; and q variable, p constant along each of the lines of the other system. Any point whatever on the surface can be regarded as the intersection of a line of the first system with a line of the second; and then the element of the first line adjacent to this point and corresponding to a variation dp will be $= \sqrt{E} \cdot dp$, and the element of the second line corresponding to the variation dq will be $= \sqrt{G} \cdot dq$. Finally, denoting by ω the angle between these elements, it is easily seen that we shall have $\cos \omega = \frac{F}{\sqrt{EG}}$. Furthermore, the area of the surface element in the form of a parallelogram between the two lines of the first system, to which correspond $q, q + dq$, and the two lines of the second system, to which correspond $p, p + dp$, will be $\sqrt{(EG - F^2)} dp \cdot dq$.

Any line whatever on the curved surface belonging to neither of the two systems is determined when p and q are supposed to be functions of a new variable, or one of them is supposed to be a function of the other. Let s be the length of such a curve, measured from an arbitrary initial point, and in either direction chosen as positive. Let θ denote the angle which the element $ds = \sqrt{(E dp^2 + 2F dp \cdot dq + G dq^2)}$ makes with the line of the first system drawn through the initial point of the element, and, in order that no ambiguity may arise, let us suppose that this angle is measured from that branch of the first line on which the values of p increase, and is taken as positive toward that side toward which the values of q increase. These conventions being made, it is easily seen that

$$\begin{aligned}\cos \theta \cdot ds &= \sqrt{E} \cdot dp + \sqrt{G} \cdot \cos \omega \cdot dq = \frac{E dp + F dq}{\sqrt{E}} \\ \sin \theta \cdot ds &= \sqrt{G} \cdot \sin \omega \cdot dq = \frac{\sqrt{(EG - F^2)} \cdot dq}{\sqrt{E}}\end{aligned}$$

18.

We shall now investigate the condition that this line be a shortest line. Since its length s is expressed by the integral

$$s = \int \sqrt{(E dp^2 + 2F dp \cdot dq + G dq^2)}$$

the condition for a minimum requires that the variation of this integral arising from an infinitely small change in the position become $= 0$. The calculation, for our purpose, is more simply made in this case, if we regard p as a function of q .

second large display on page 105 (after the words “Thus we have”). For a curve parameterized by arclength, this equation says that

$$\begin{aligned} \frac{\partial E}{\partial p}(c(s)) \left(\frac{dc^1}{ds} \right)^2 + 2 \frac{\partial F}{\partial p}(c(s)) \frac{dc^1}{ds} \frac{dc^2}{ds} + \frac{\partial G}{\partial p}(c(s)) \left(\frac{dc^2}{ds} \right)^2 \\ = 2 \frac{d}{ds} \left[E(c(s)) \frac{dc^1}{ds} + F(c(s)) \frac{dc^2}{ds} \right]. \end{aligned}$$

It is a very useful exercise to write out the equations on pg. I.329 for the case of a 2-dimensional manifold, with $g_{11} = E$, $g_{12} = F$, $g_{22} = G$, and show that the first of these equations (the equation for $k = 1$) yields the above equation (it will be necessary to perform the differentiation on the right side).

Although Gauss performs various further manipulations, it is only necessary to follow the next step,

$$2 \frac{d}{ds} \left[E \frac{dc^1}{ds} + F \frac{dc^2}{ds} \right] = 2 \frac{d}{ds} \sqrt{E} \cos \theta,$$

where θ is defined in the previous section.

§19. In this section Gauss rewrites formulas from preceding sections for the case of a coordinate system (p, q) which is “orthogonal” ($\langle \partial/\partial p, \partial/\partial q \rangle = F = 0$). The important case for us is the last he considers, in which the coordinates are

When this is done, if the variation is denoted by the characteristic δ , we have

$$\begin{aligned}\delta s &= \int \frac{\left(\frac{dE}{dp} \cdot dp^2 + \frac{2dF}{dp} \cdot dp \cdot dq + \frac{dG}{dp} \cdot dq^2\right) \delta p + (2E dp + 2F dq) d \delta p}{2 ds} \\ &= \frac{E dp + F dq}{ds} \cdot \delta p + \int \delta p \cdot \left\{ \frac{\frac{dE}{dp} \cdot dp^2 + \frac{2dF}{dp} \cdot dp \cdot dq + \frac{dG}{dp} \cdot dq^2}{2 ds} - d \cdot \frac{E dp + F dq}{ds} \right\}\end{aligned}$$

and we know that what is included under the integral sign must vanish independently of δp . Thus we have

$$\begin{aligned}\frac{dE}{dp} \cdot dp^2 + \frac{2dF}{dp} \cdot dp \cdot dq + \frac{dG}{dp} \cdot dq^2 &= 2 ds \cdot d \cdot \frac{E dp + F dq}{ds} \\ &= 2 ds \cdot d \cdot \sqrt{E} \cdot \cos \theta = \frac{ds \cdot dE \cdot \cos \theta}{\sqrt{E}} - 2 ds \cdot d\theta \cdot \sqrt{E} \cdot \sin \theta \\ &= \frac{(E dp + F dq) dE}{E} - 2\sqrt{(EG - F^2)} \cdot dq \cdot d\theta \\ &= \left(\frac{E dp + F dq}{E}\right) \cdot \left(\frac{dE}{dp} \cdot dp + \frac{dE}{dq} \cdot dq\right) - 2\sqrt{(EG - F^2)} \cdot dq \cdot d\theta\end{aligned}$$

This gives the following conditional equation for a shortest line:

$$\sqrt{(EG - F^2)} \cdot d\theta = \frac{1}{2} \cdot \frac{F}{E} \cdot \frac{dE}{dp} \cdot dp + \frac{1}{2} \cdot \frac{F}{E} \cdot \frac{dE}{dq} \cdot dq + \frac{1}{2} \cdot \frac{dE}{dq} \cdot dp - \frac{dF}{dp} \cdot dp - \frac{1}{2} \cdot \frac{dG}{dp} \cdot dq$$

which can also be written

$$\sqrt{(EG - F^2)} \cdot d\theta = \frac{1}{2} \cdot \frac{F}{E} \cdot dE + \frac{1}{2} \cdot \frac{dE}{dq} \cdot dp - \frac{dF}{dp} \cdot dp - \frac{1}{2} \cdot \frac{dG}{dp} \cdot dq$$

From this equation, by means of the equation

$$\cot \theta = \frac{E}{\sqrt{(EG - F^2)}} \cdot \frac{dp}{dq} + \frac{F}{\sqrt{(EG - F^2)}}$$

it is also possible to eliminate the angle θ , and to derive a differential equation of the second order between p and q , which, however, would become more complicated and less useful for applications than the preceding.

The general formulæ, which we have derived in Arts. 11, 18 for the measure of curvature and the variation in the direction of a shortest line, become much simpler if the quantities p, q are so chosen that the lines of the first system cut

the “polar coordinates” (r, ϕ) defined in terms of the geodesics emanating from a point A of the surface. Here Gauss obtains the formula

$$k = -\frac{1}{\sqrt{G}} \frac{\partial^2 \sqrt{G}}{\partial r^2},$$

and

$$\frac{d\theta}{ds} = -\frac{\partial \sqrt{G}}{\partial r} \frac{d\phi(c(s))}{ds},$$

where θ is the angle the geodesic c makes with the lines $\phi = \text{constant}$. Notice that (r, ϕ) is not a coordinate system on a whole neighborhood of A ; we must delete one geodesic ray, including the point A itself. Consequently, \sqrt{G} and $\partial \sqrt{G} / \partial r$ are not even defined at A . Gauss’ final assertions in this section should be interpreted as saying that

$$\begin{aligned} \lim_{B \rightarrow A} \sqrt{G}(B) &= 0 \\ \lim_{B \rightarrow A} \frac{\partial \sqrt{G}}{\partial r}(B) &= 1. \end{aligned}$$

everywhere orthogonally the lines of the second system; i.e., in such a way that we have generally $\omega = 90^\circ$, or $F = 0$. Then the formula for the measure of curvature becomes

$$4E^2G^2k = E \cdot \frac{dE}{dq} \cdot \frac{dG}{dq} + E \left(\frac{dG}{dp} \right)^2 + G \cdot \frac{dE}{dp} \cdot \frac{dG}{dp} + G \left(\frac{dE}{dq} \right)^2 - 2EG \left(\frac{ddE}{dq^2} + \frac{ddG}{dp^2} \right)$$

and for the variation of the angle θ

$$\sqrt{EG} \cdot d\theta = \frac{1}{2} \cdot \frac{dE}{dq} \cdot dp - \frac{1}{2} \cdot \frac{dG}{dp} \cdot dq$$

Among the various cases in which we have this condition of orthogonality, the most important is that in which all the lines of one of the two systems, e.g., the first, are shortest lines. Here for a constant value of q the angle θ becomes $= 0$, and therefore the equation for the variation of θ just given shows that we must have $\frac{dE}{dq} = 0$, or that the coefficient E must be independent of q ; i.e., E must be either a constant or a function of p alone. It will be simplest to take for p the length of each line of the first system, which length, when all the lines of the first system meet in a point, is to be measured from this point, or, if there is no common intersection, from any line whatever of the second system. Having made these conventions, it is evident that p and q denote now the same quantities that were expressed in Arts. 15, 16 by r and φ , and that $E = 1$. Thus the two preceding formulæ become:

$$4G^2k = \left(\frac{dG}{dp} \right)^2 - 2G \frac{ddG}{dp^2}$$

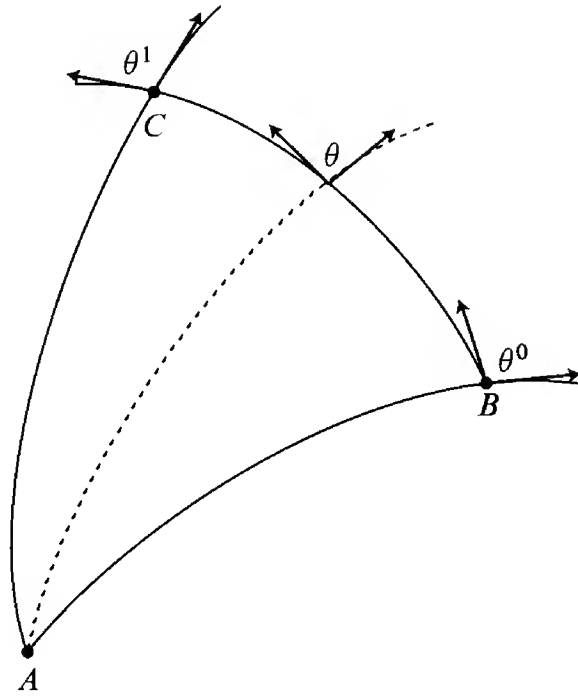
$$\sqrt{G} \cdot d\theta = -\frac{1}{2} \cdot \frac{dG}{dp} \cdot dq$$

or, setting $\sqrt{G} = m$,

$$k = -\frac{1}{m} \cdot \frac{ddm}{dp^2}, \quad d\theta = -\frac{dm}{dp} \cdot dq$$

Generally speaking, m will be a function of p , q , and $m dq$ the expression for the element of any line whatever of the second system. But in the particular case where all the lines p go out from the same point, evidently we must have $m = 0$ for $p = 0$. Furthermore, in the case under discussion we will take for q the angle itself which the first element of any line whatever of the first system makes with the element of any one of the lines chosen arbitrarily. Then, since for an infinitely small value of p the element of a line of the second system (which can be regarded as a circle described with radius p) is $= p dq$, we shall have for

§20. If you have come this far, there should be no problem with this final section. Here is the picture.



an infinitely small value of p , $m = p$, and consequently, for $p = 0$, $m = 0$ at the same time, and $\frac{dm}{dp} = 1$.

20.

We pause to investigate the case in which we suppose that p denotes in a general manner the length of the shortest line drawn from a fixed point A to any other point whatever of the surface, and q the angle that the first element of this line makes with the first element of another given shortest line going out from A . Let B be a definite point in the latter line, for which $q = 0$, and C another definite point of the surface, at which we denote the value of q simply by A . Let us suppose the points B, C joined by a shortest line, the parts of which, measured from B , we denote in a general way, as in Art. 18, by s ; and, as in the same article, let us denote by θ the angle which any element ds makes with the element dp ; finally, let us denote by θ^0, θ' the values of the angle θ at the points B, C . We have thus on the curved surface a triangle formed by shortest lines. The angles of this triangle at B and C we shall denote simply by the same letters, and B will be equal to $180^0 - \theta$, C to θ' itself. But, since it is easily seen from our analysis that all the angles are supposed to be expressed, not in degrees, but by numbers, in such a way that the angle $57^0 17' 45''$, to which corresponds an arc equal to the radius, is taken for the unit, we must set

$$\theta^0 = \pi - B, \quad \theta' = C$$

where 2π denotes the circumference of the sphere. Let us now examine the integral curvature of this triangle, which is $= \int k d\sigma$, $d\sigma$ denoting a surface element of the triangle. Wherefore, since this element is expressed by $m dp \cdot dq$, we must extend the integral $\iint m dp \cdot dq$ over the whole surface of the triangle. Let us begin by integration with respect to p , which, because $k = -\frac{1}{m} \cdot \frac{ddm}{dp^2}$, gives $dq \cdot (\text{Const.} - \frac{dm}{dp})$, for the integral curvature of the area lying between the lines of the first system, to which correspond the values $q, q + dq$ of the second indeterminate. Since this integral curvature must vanish for $p = 0$, the constant introduced by integration must be equal to the value of $\frac{dm}{dq}$ for $p = 0$, i.e., equal to unity. Thus we have $dq(1 - \frac{dm}{dp})$, where for $\frac{dm}{dp}$ must be taken the value corresponding to the end of this area on the line CB . But on this line we have, by the preceding article, $\frac{dm}{dq} \cdot dq = -d\theta$, whence our expression is changed into $dq + d\theta$. Now by a second integration, taken from $q = 0$ to $q = A$, we find that the integral curvature $= A + \theta' - \theta^0 = A + B + C - \pi$.

The integral curvature is equal to the area of that part of the sphere which corresponds to the triangle, taken with the positive or negative sign according as the curved surface on which the triangle lies is concavo-concave or concavo-convex. For unit area will be taken the square whose side is equal to unity (the radius of the sphere), and then the whole surface of the sphere becomes $= 4\pi$. Thus the part of the surface of the sphere corresponding to the triangle is to the whole surface of the sphere as $\pm(A + B + C - \pi)$ is to 4π . This theorem, which, if we mistake not, ought to be counted among the most elegant in the theory of curved surfaces, may also be stated as follows:

The excess over 180^0 of the sum of the angles of a triangle formed by shortest lines on a concavo-concave curved surface, or the deficit from 180^0 of the sum of the angles of a triangle formed by shortest lines on a concavo-convex curved surface, is measured by the area of the part of the sphere which corresponds, through the directions of the normals, to that triangle, if the whole surface of the sphere is set equal to 720 degrees.

More generally, in any polygon whatever of n sides, each formed by a shortest line, the excess of the sum of the angles over $(2n - 4)$ right angles, or the deficit from $(2n - 4)$ right angles (according to the nature of the curved surface), is equal to the area of the corresponding polygon on the sphere, if the whole surface of the sphere is set equal to 720 degrees. This follows at once from the preceding theorem by dividing the polygon into triangles.

B. GAUSS' THEORY OF SURFACES

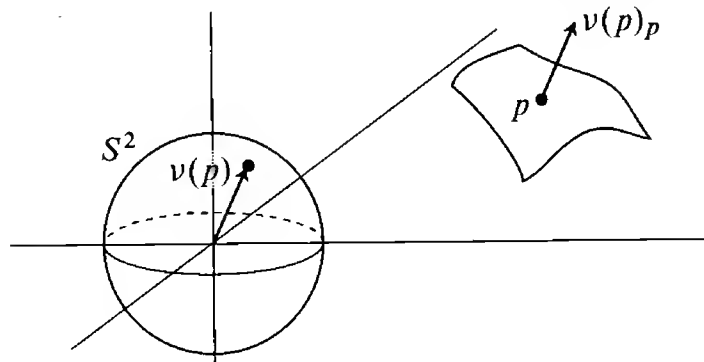
This part of the chapter presents Gauss' results in modern dressing. It can be read completely independently of the first part, but there are frequent comparisons with Gauss' original paper. Some of Gauss' notation will be changed; in particular, we will use K for Gauss' k .

We consider a 2-dimensional submanifold M of \mathbb{R}^3 , with $i: M \rightarrow \mathbb{R}^3$ the inclusion map. We also assume that M has been oriented. Since all our results will be local ones, we merely need an orientation in a neighborhood of each point, so this assumption does not place any real restriction on M . (However, we must still investigate to what extent our results depend on the choice of the orientation.)

At each point $p \in M$ there is a unique unit vector $v(p) \in \mathbb{R}^3$ such that

- (1) $v(p)_p \in \mathbb{R}^3_p$ is perpendicular to M_p
- (2) $v(p), v, w$ is positively oriented in \mathbb{R}^3 whenever $v_p, w_p \in M_p$ is positively oriented.

We thus have the **normal map** $v: M \rightarrow \mathbb{R}^3$, which actually goes to the unit sphere, $v: M \rightarrow S^2 \subset \mathbb{R}^3$. Notice that in his paper Gauss uses X, Y, Z for



$v^1(p), v^2(p), v^3(p)$. The idea of using this map may have been suggested to Gauss by astronomical practices, as he indicates in an abstract of the paper (also included in the Princeton University Library translation, and the Raven Press reprint). At any rate, the map v turns out to play such a crucial role that it is often called the **Gauss map**.

An explicit formula for $v: M \rightarrow S^2$ can be obtained from various explicit descriptions of M . For example, if $M = \{p \in \mathbb{R}^3 : W(p) = 0\}$ for some function $W: \mathbb{R}^3 \rightarrow \mathbb{R}$, then $dW = 0$ on M , i.e., $dW(v_q) = 0$ for all $v_q \in M_q$. This means that

$$D_1 W(q) \cdot v^1 + D_2 W(q) \cdot v^2 + D_3 W(q) \cdot v^3 = 0$$

$$\frac{\partial W}{\partial x}(q) \cdot v^1 + \frac{\partial W}{\partial y}(q) \cdot v^2 + \frac{\partial W}{\partial z}(q) \cdot v^3 = 0 \quad \text{for all } v_q \in M_q.$$

This equation can be written

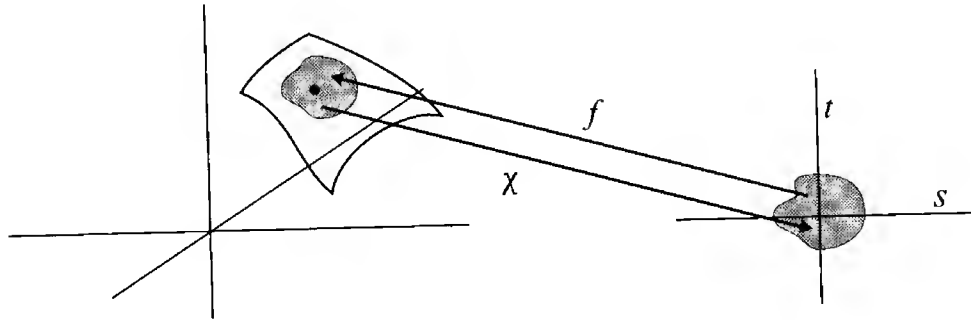
$$\left\langle \left(\frac{\partial W}{\partial x}(q), \frac{\partial W}{\partial y}(q), \frac{\partial W}{\partial z}(q) \right), v \right\rangle = 0 \quad \text{for all } v_q \in M_q.$$

Consequently,

$$v(q) = \text{normalized} \left(\frac{\partial W}{\partial x}(q), \frac{\partial W}{\partial y}(q), \frac{\partial W}{\partial z}(q) \right),$$

which is precisely the formula Gauss gives.

We can also find v in terms of a coordinate system χ . To avoid confusion, we will denote the standard coordinate system in \mathbb{R}^2 by (s, t) [Gauss uses (p, q)]



and we will denote the inverse function $\chi^{-1}: \mathbb{R}^2 \rightarrow M \subset \mathbb{R}^3$ by f . It is naturally necessary to consider the component functions of f , considered as a map into \mathbb{R}^3 , in order to obtain a formula for v ; we cannot obtain a formula for v totally in terms of χ , since this coordinate system tells us nothing about the way M is situated in \mathbb{R}^3 . Note that if $q = f(s, t)$, then

$$\begin{aligned} \frac{\partial}{\partial \chi^1} \Big|_q &= \left(\frac{\partial f}{\partial s}(s, t) \right)_q = \left(\frac{\partial f^1}{\partial s}(s, t), \frac{\partial f^2}{\partial s}(s, t), \frac{\partial f^3}{\partial s}(s, t) \right)_q \\ \frac{\partial}{\partial \chi^2} \Big|_q &= \left(\frac{\partial f}{\partial t}(s, t) \right)_q = \left(\frac{\partial f^1}{\partial t}(s, t), \frac{\partial f^2}{\partial t}(s, t), \frac{\partial f^3}{\partial t}(s, t) \right)_q. \end{aligned}$$

Consequently,

$$v(f(s, t)) = \text{normalized cross product} \left(\frac{\partial f^1}{\partial s}, \frac{\partial f^2}{\partial s}, \frac{\partial f^3}{\partial s} \right) \times \left(\frac{\partial f^1}{\partial t}, \frac{\partial f^2}{\partial t}, \frac{\partial f^3}{\partial t} \right).$$

Thus we have

$$\nu(f(s, t)) = \pm \frac{\frac{\partial f^2}{\partial s} \frac{\partial f^3}{\partial t} - \frac{\partial f^2}{\partial t} \frac{\partial f^3}{\partial s}}{\Delta}, \quad \text{---}, \quad \text{---},$$

for

$\Delta = \text{norm of the cross product,}$

exactly as in Gauss.

Finally, if M is the graph of $g: \mathbb{R}^2 \rightarrow \mathbb{R}$, so that

$$M = \{(x, y, g(x, y)) : x, y \in \mathbb{R}^2\},$$

then M is the image of $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by

$$f(s, t) = (s, t, g(s, t)).$$

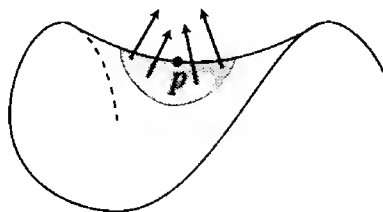
It follows that

$$\nu(x, y, g(x, y)) = \text{normalized cross product} \left(1, 0, \frac{\partial g}{\partial x}\right) \times \left(0, 1, \frac{\partial g}{\partial y}\right).$$

We are now ready for a preliminary, non-rigorous, definition of the **curvature** $K(p)$ of M at p :

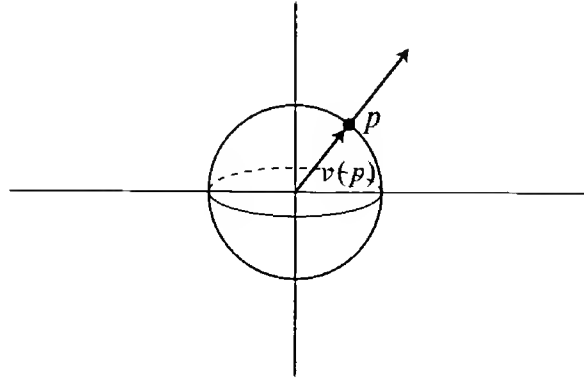
$$K(p) = \lim_{A \rightarrow p} \frac{\text{area } \nu(A)}{\text{area } A},$$

where the limit is taken as the region A around p becomes smaller and smaller. There would be considerable difficulties involved in making this definition rigorous. In the first place, we would have to prove that the limit exists. More crucial, the “area of $\nu(A)$ ” needs some interpretation; in the figure below we want “area of $\nu(A)$ ” to be negative, because ν is orientation reversing near p .



However, even with this non-rigorous definition we can find the curvature of certain surfaces.

Consider first the surface S^2 . The map $v: S^2 \rightarrow S^2$ is just the identity, so at



each point $p \in S^2$ we have

$$K(p) = \lim_{A \rightarrow p} \frac{\text{area } v(A)}{\text{area } A} = \lim_{A \rightarrow p} \frac{\text{area } A}{\text{area } A} = 1.$$

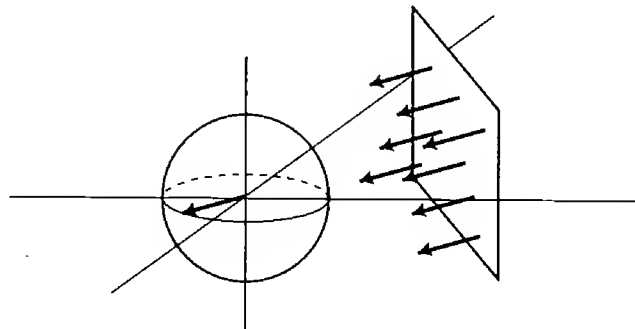
For the case of the sphere $S^2(r)$ of radius r , we have

$$\begin{aligned} K(p) &= \lim_{A \rightarrow p} \frac{\text{area } v(A)}{\text{area } A} \\ &= \frac{1/r^2 \cdot \text{area } A}{\text{area } A} \\ &= \frac{1}{r^2}, \end{aligned}$$

Two diagrams of spheres. The left sphere is labeled $S^2(r)$ and has a shaded region A on its surface. The right sphere is labeled S^2 and has a smaller shaded region $v(A)$ on its surface. Arrows indicate the mapping from A to $v(A)$.

which certainly seems reasonable.

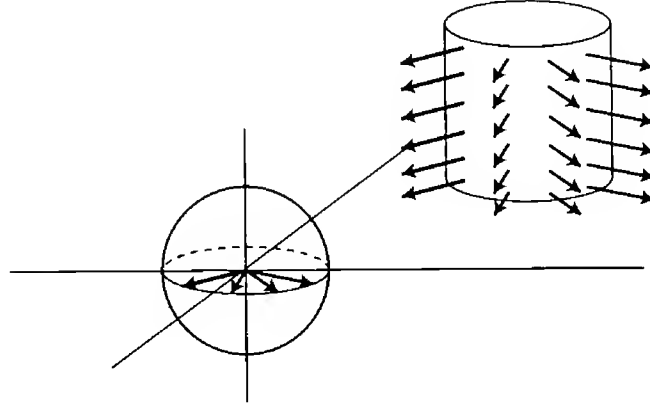
Next we consider a plane P . The function $v: P \rightarrow S^2$ is constant, so



$$K(p) = \lim_{A \rightarrow p} \frac{\text{area } v(A)}{\text{area } A} = \lim_{A \rightarrow p} \frac{0}{\text{area } A} = 0;$$

the plane does not curve.

Finally, we consider a cylinder Z . In this case, the function $v: Z \rightarrow S^2$ is not constant, but the image of v always lies along a certain arc in S^2 , so for all



points $p \in Z$ we have

$$K(p) = \lim_{A \rightarrow p} \frac{\text{area } v(A)}{\text{area } A} = \lim_{A \rightarrow p} \frac{0}{\text{area } A} = 0.$$

The cylinder, too, does not curve! It begins to look as if we have the “wrong” definition of curvature; only later will Gauss explain why this is the “right” definition.

The manner in which we produce a rigorous definition of curvature is really very simple. Since M is a submanifold of \mathbb{R}^3 , which has the usual Riemannian metric $\langle \cdot, \cdot \rangle$, we can give M the induced Riemannian metric $i^*\langle \cdot, \cdot \rangle$. Together with the orientation which we have given M , this metric determines a 2-form “ dV ” on M , namely

$$dV(q)(v_q, w_q) = \text{signed area of the parallelogram spanned by } v \text{ and } w.$$

On the sphere S^2 we also have a volume element, coming from its induced Riemannian metric and its usual orientation. As a glance at pg. I.264 will show, this is just the 2-form which we have denoted by σ' .

We now define the **Gaussian curvature** $K(p)$ of M at p to be

$$K(p) = \frac{v^*(\sigma')(p)}{dV(p)};$$

in this equation, the division of 2-forms makes sense because any 2-form on the 2-dimensional manifold M is a multiple of the non-zero 2-form dV . If the

vectors $v_p, w_p \in M_p$ are orthonormal, then our definition says that

$$\begin{aligned} K(p) &= v^*(\sigma')(p)(v_p, w_p) \\ &= \sigma'(v(p))(v_*v_p, v_*w_p) \quad \text{for orthonormal } v_p, w_p. \end{aligned}$$

This definition of $K(p)$ involves v , and hence the orientation μ which we picked for a neighborhood of p . Choosing the opposite orientation $-\mu$ changes v to $-v = A \circ v$, where $A: S^2 \rightarrow S^2$ is the antipodal map, and consequently changes

$$v^*(\sigma') \quad \text{to} \quad v^*(A^*(\sigma')) = -v^*(\sigma').$$

On the other hand, dV is also changed to $-dV$, so $K(p)$ does not depend on the choice of orientation.

Notice that if v is one-one in a neighborhood of p , then for every region A contained in that neighborhood we have

$$\begin{aligned} \text{area } v(A) &= \int_{v(A)} \sigma' \\ &= \pm \int_A v^*(\sigma') \quad \text{depending on whether } v \text{ is} \\ &\quad \text{orientation preserving or reversing.} \end{aligned}$$

Consequently,

$$\frac{\text{area } v(A)}{\text{area } A} = \frac{\pm \int_A v^*(\sigma')}{\int_A dV}.$$

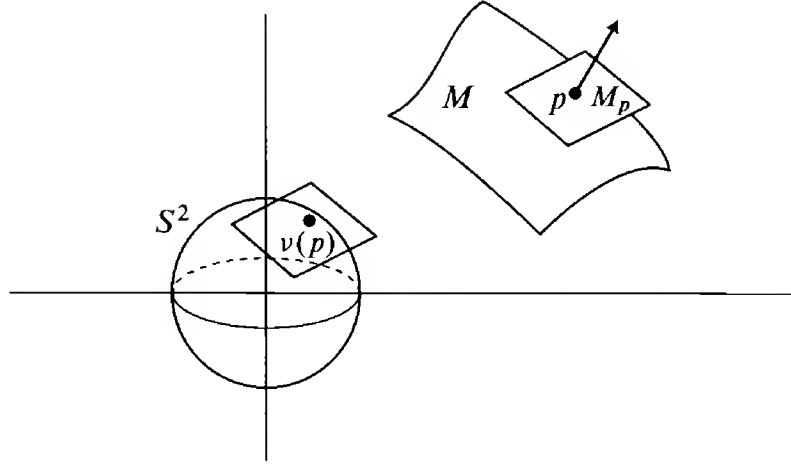
We thus recover our original “definition”, provided that

$$\lim_{A \rightarrow p} \frac{\pm \int_A v^*(\sigma')}{\int_A dV} = \pm \frac{v^*(\sigma')(p)}{dV(p)} = \pm K(p);$$

although this result seems reasonable, by continuity of $v^*(\sigma')$ and dV , we will not worry about the exact manner in which A must approach p in order for the limit to work out. If v is not one-one in a neighborhood of p , then we must have $v_*(p) = 0$, so the rigorous definition gives $K(p) = 0$, which is more or less what one would expect from the intuitive definition.

Although we will eventually obtain a neater expression for K , we begin by deriving Gauss’ first formula for K , using essentially the same reasoning as

Gauss uses. We first observe that the tangent plane M_p is *parallel* to the tangent plane $S^2_{v(p)}$ of S^2 at $v(p)$. The reason is very simple: the tangent plane $S^2_{v(p)}$



is perpendicular to $v(p)$, and so is M_p , by the very definition of $v(p)$. In the previous two sentences we have used the identification of \mathbb{R}^3_p with \mathbb{R}^3 , the identification of M_p with $i_*M_p \subset \mathbb{R}^3_p$, etc. Without further warning, we shall continue to do so, to avoid cluttering up the page with extra symbolism.

If $v_p, w_p \in M_p$ are linearly independent, then

$$K(p) = \frac{\text{area of parallelogram } P \text{ spanned by } v_*v_p, v_*w_p}{\text{area of parallelogram } Q \text{ spanned by } v_p, w_p}.$$

Since M_p is parallel to $S^2_{v(p)}$, this implies that

$$K(p) = \frac{\text{area of projection of } P \text{ on } (x, y)\text{-plane}}{\text{area of projection of } Q \text{ on } (x, y)\text{-plane}}$$

(provided the denominator is not zero). In particular, suppose M is the graph of $g: \mathbb{R}^2 \rightarrow \mathbb{R}$, and consequently the image of $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by

$$f(s, t) = (s, t, g(s, t)).$$

We choose

$$v = \frac{\partial f}{\partial s} = \left(1, 0, \frac{\partial g}{\partial s}\right)$$

$$w = \frac{\partial f}{\partial t} = \left(0, 1, \frac{\partial g}{\partial t}\right)$$

and consider $v_p, w_p \in M_p$, for $p = f(s, t)$. Then

$$\begin{aligned} & \text{area of projection of } Q \text{ on } (x, y)\text{-plane} \\ &= \text{area of parallelogram spanned by } (1, 0) \text{ and } (0, 1) \\ &= 1. \end{aligned}$$

On the other hand,

$$\begin{aligned} v_*(v_p) &= v_* \left(\left. \frac{\partial f}{\partial s} \right|_p \right) = v_* f_* \left(\frac{\partial}{\partial s} \right) \\ &= (v \circ f)_* \left(\frac{\partial}{\partial s} \right) \\ &= \left(\frac{\partial(v \circ f)}{\partial s} \right)_{v(p)} = \left(\frac{\partial(v^1 \circ f)}{\partial s}, \frac{\partial(v^2 \circ f)}{\partial s}, \frac{\partial(v^3 \circ f)}{\partial s} \right)_{v(p)}. \end{aligned}$$

(Here all partial derivatives are to be evaluated at the point (s, t) .) Similarly,

$$v_*(w_p) = \left(\frac{\partial(v^1 \circ f)}{\partial t}, \frac{\partial(v^2 \circ f)}{\partial t}, \frac{\partial(v^3 \circ f)}{\partial t} \right)_{v(p)}.$$

Consequently,

$$\begin{aligned} K(f(s, t)) &= \text{area of projection of } P \text{ on } (x, y)\text{-plane} \\ &= \text{area of parallelogram spanned by} \\ &\quad \left(\frac{\partial(v^1 \circ f)}{\partial s}, \frac{\partial(v^2 \circ f)}{\partial s} \right) \text{ and } \left(\frac{\partial(v^1 \circ f)}{\partial t}, \frac{\partial(v^2 \circ f)}{\partial t} \right) \\ &= \frac{\partial(v^1 \circ f)}{\partial s} \frac{\partial(v^2 \circ f)}{\partial t} - \frac{\partial(v^1 \circ f)}{\partial t} \frac{\partial(v^2 \circ f)}{\partial s}. \end{aligned}$$

This is precisely the formula Gauss obtains, at the top of page 77. If we use the formula

$$v(x, y, g(x, y)) = \text{normalized cross product } \left(1, 0, \frac{\partial g}{\partial x} \right) \times \left(0, 1, \frac{\partial g}{\partial y} \right),$$

we obtain, after a little calculation,

$$(*) \quad K(x, y, g(x, y)) = \frac{\frac{\partial^2 g}{\partial x^2} \frac{\partial^2 g}{\partial y^2} - \left(\frac{\partial^2 g}{\partial x \partial y} \right)^2}{\left(1 + \left(\frac{\partial g}{\partial x} \right)^2 + \left(\frac{\partial g}{\partial y} \right)^2 \right)^{3/2}}.$$

We can compare this with the results of Chapter 2, in which we picked our coordinate system so that

$$p = (0, 0, 0) = (0, 0, g(0, 0))$$

$$\frac{\partial g}{\partial x}(0, 0) = \frac{\partial g}{\partial y}(0, 0) = 0$$

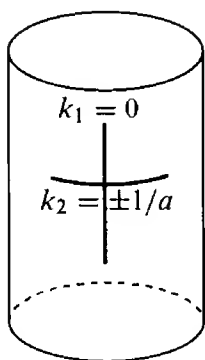
$$\frac{\partial^2 g}{\partial x \partial y}(0, 0) = 0.$$

In this case, we obtained the result that the minimum k_1 and maximum k_2 of all curvatures cut out by normal planes through p are the minimum and maximum of $\partial^2 g / \partial x^2(0, 0)$ and $\partial^2 g / \partial y^2(0, 0)$. Now for our special choice of coordinates in \mathbb{R}^3 , formula (*) becomes

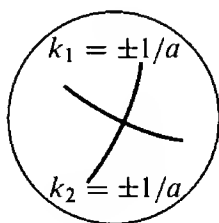
$$K(p) = \frac{\partial^2 g}{\partial x^2} \frac{\partial^2 g}{\partial y^2} = k_1 \cdot k_2.$$

(Notice that, just as K does not depend on the orientation of M , neither does the *product* $k_1 \cdot k_2$, even though k_1 and k_2 individually do). We thus have the following result:

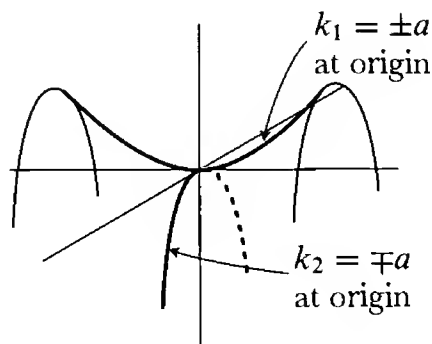
The Gaussian curvature $K(p)$ at any point $p \in M$ is the product of the extreme curvatures of the curves through p cut out by normal planes.



cylinder of radius a
 $K = k_1 \cdot k_2 = 0$



sphere of radius a
 $K = k_1 \cdot k_2 = \frac{1}{a^2}$



graph of $z = \frac{ax^2}{2} - \frac{ay^2}{2}$
 $K(0, 0, 0) = k_1 \cdot k_2 = -a^2$

To prove this result we have followed Gauss' exposition. In particular, the proof of Euler's Theorem which appeared in Chapter 2 is Gauss'. Undoubtedly, this proof is considerably simpler than Euler's, for Gauss remarks with pride

“These conclusions contain almost all that the illustrious Euler was the first to prove on the curvature of curved surfaces.” Nevertheless, later developments provided a nicer way of obtaining these results, which will consequently now be rederived.

The definition of curvature involves the map $\nu: M \rightarrow S^2 \subset \mathbb{R}^3$, but even more important, it involves the map $\nu_*: M_p \rightarrow S^2_{\nu(p)}$. We can also think of ν as an \mathbb{R}^3 -valued function $\nu: M \rightarrow \mathbb{R}^3$, and we then have a map

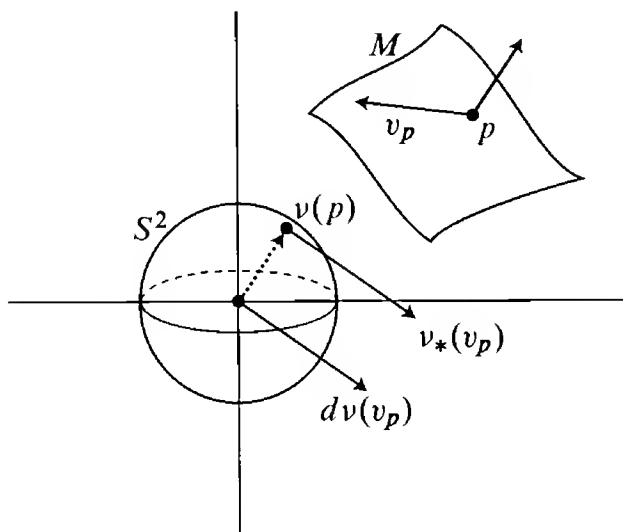
$$d\nu: M_p \rightarrow \mathbb{R}^3,$$

namely

$$\begin{aligned} d\nu(v_p) &= (dv^1(v_p), dv^2(v_p), dv^3(v_p)) \\ &= (v_p(v^1), v_p(v^2), v_p(v^3)). \end{aligned}$$

We claim that $d\nu$ is essentially the same as ν_* ; to be precise,

$$\nu_*(v_p) = d\nu(v_p)_{\nu(p)}.$$



Probably the easiest way* to see this is to take a curve c in M with $c'(0) = v_p$. Then

$$\begin{aligned} \nu_*(v_p) &= (\nu \circ c)'(0) \\ &= \left(\frac{d(v^1 \circ c)}{dt} \Big|_{t=0}, \frac{d(v^2 \circ c)}{dt} \Big|_{t=0}, \frac{d(v^3 \circ c)}{dt} \Big|_{t=0} \right)_{\nu(p)} \\ &= (v_p(v^1), v_p(v^2), v_p(v^3))_{\nu(p)}. \end{aligned}$$

*A more formal way is to introduce the inclusion map $j: S^2 \rightarrow \mathbb{R}^3$, and observe that we are really trying to prove that

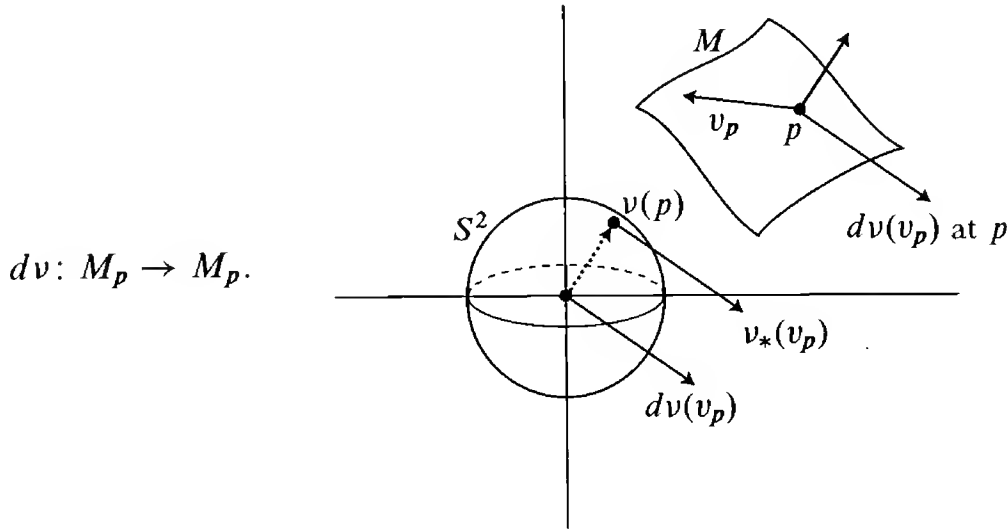
$$j_*\nu_*(v_p) = d(j \circ \nu)_{\nu(p)}.$$

The result then follows from Problem I.4-3, applied to the component functions of ν .

Taking advantage of the identification of \mathbb{R}^3 with \mathbb{R}^3_p , we can introduce one more confusion, and consider dv as a map

$$dv: M_p \rightarrow \mathbb{R}^3_p.$$

Since $dv(v_p)$ is parallel to $v_*(v_p) \in S^2_{v(p)}$, and since $S^2_{v(p)}$ is parallel to M_p , we see that we actually have a map



Despite the tortuous process used to define this map, the net result can be described very simply: $dv(v_p)$ is just $v_*(v_p)$ moved back up to a parallel vector in M_p .

The map $dv: M_p \rightarrow M_p$ is sometimes called the **Weingarten map**. Using it, we can define a tensor \mathbf{II} on M which is covariant of order 2: for $v_p, w_p \in M_p$ we define

$$\mathbf{II}(p)(v_p, w_p) = -\langle dv(v_p), w_p \rangle.$$

Notice that \mathbf{II} *does* depend on the choice of v , and hence on the orientation picked for M . This tensor \mathbf{II} is called the **second fundamental form** of M . But, as we shall soon see, it is *not* alternating, and hence not a 2-form. We can see the reason for the word “form” in the classical terminology by looking at another tensor that you may be worrying about. The **first fundamental form** \mathbf{I} of M is just the induced Riemannian metric $i^*(\langle \cdot, \cdot \rangle)$ on M (where $\langle \cdot, \cdot \rangle$ is the usual Riemannian metric on \mathbb{R}^3). Thus

$$\mathbf{I}(p)(v_p, w_p) = \langle v_p, w_p \rangle \quad [= \langle v, w \rangle].$$

Classically, one worked not with \mathbf{I} and \mathbf{II} but with the functions $v \mapsto \mathbf{I}(v, v)$ and $v \mapsto \mathbf{II}(v, v)$, which are “quadratic forms” in the components of v (compare pg. I.314).

The second fundamental form provides us with a name for a quantity that appeared in Chapter 2:

0. PROPOSITION. Let c be a curve in M which is parameterized by arc-length. Let $c(0) = p$, and let $X = c'(0) \in M_p$. Then

$$\langle c''(0), \nu(p) \rangle = \Pi(X, X) \quad [\text{i.e., } \Pi(p)(X, X)].$$

Consequently, $\Pi(X, X)$ is the signed curvature κ_X of the curve cut out on M by the normal plane through $\nu(p)$ and X (with $X, \nu(p)$ positively oriented). Moreover, if κ_ϕ is the curvature of the curve c_ϕ cut out by the plane which contains X and makes an angle of ϕ with the normal plane, then

$$\kappa_\phi \cdot \cos \phi = \kappa_X.$$

PROOF. Clearly

$$(1) \quad \left. \frac{d\nu(c(s))}{ds} \right|_{s=0} = d\nu(X).$$

Since

$$\langle c'(s), \nu(c(s)) \rangle = 0 \quad \text{for all } s,$$

differentiation yields, using (1),

$$\begin{aligned} \langle c''(0), \nu(p) \rangle &= -\langle c'(0), d\nu(X) \rangle \\ &= -\langle X, d\nu(X) \rangle \\ &= \Pi(X, X). \end{aligned}$$

This, of course, is precisely the result (proved in precisely the same way) which was derived in Chapter 2. The rest of the theorem is proved just as before. ♦

Unlike Meusnier's Theorem, which involves Π in a trivial way, Euler's Theorem involves a crucial property of Π :

1. THEOREM. The second fundamental form Π of M is symmetric,

$$\Pi(p)(X_p, Y_p) = \Pi(p)(Y_p, X_p) \quad \text{for } X_p, Y_p \in M_p.$$

FIRST PROOF. Near p we can represent M as the image of a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$. Let $N = \nu \circ f$, so that N is " ν , considered as a function on \mathbb{R}^2 ". Then

$$d\nu \left(\frac{\partial f}{\partial s} \right) = d\nu \left(f_* \left(\frac{\partial}{\partial s} \right) \right) = f_* \left(\frac{\partial}{\partial s} \right) (\nu) = \frac{\partial}{\partial s} (\nu \circ f) = \frac{\partial N}{\partial s}.$$

Now we clearly have

$$\left\langle N, \frac{\partial f}{\partial t} \right\rangle = 0.$$

Differentiating with respect to s gives

$$\left\langle N, \frac{\partial^2 f}{\partial s \partial t} \right\rangle = - \left\langle \frac{\partial N}{\partial s}, \frac{\partial f}{\partial t} \right\rangle = - \left\langle dv \left(\frac{\partial f}{\partial s} \right), \frac{\partial f}{\partial t} \right\rangle = \Pi \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial t} \right).$$

Exactly the same argument gives

$$(a) \quad \left\langle N, \frac{\partial^2 f}{\partial t \partial s} \right\rangle = \Pi \left(\frac{\partial f}{\partial t}, \frac{\partial f}{\partial s} \right).$$

Thus

$$\Pi \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial t} \right) = \Pi \left(\frac{\partial f}{\partial t}, \frac{\partial f}{\partial s} \right).$$

Since $\partial f/\partial s, \partial f/\partial t$ are a basis for the tangent space of M at each point, Π is symmetric. For later use we also note that similar arguments lead to the equations

$$(b) \quad \Pi \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial s} \right) = \left\langle N, \frac{\partial^2 f}{\partial s^2} \right\rangle$$

$$(c) \quad \Pi \left(\frac{\partial f}{\partial t}, \frac{\partial f}{\partial t} \right) = \left\langle N, \frac{\partial^2 f}{\partial t^2} \right\rangle.$$

Equations (a)–(c) are called the *Weingarten equations*.

SECOND (COORDINATE FREE, FANCY) PROOF. Let Y be a vector field in a neighborhood $U \subset M$ of p whose value at p is Y_p , and such that $Y(q) \in M_q$ for all $q \in U$. Since

$$\langle v, Y \rangle = 0 \quad \text{on } U,$$

we have

$$\begin{aligned} 0 &= X_p(\langle v, Y \rangle) = \langle X_p(v), Y_p \rangle + \langle v(p), X_p(Y) \rangle \\ &= \langle dv(X_p), Y_p \rangle + \langle v(p), X_p(Y) \rangle, \end{aligned}$$

where $X_p(Y)$ denotes the vector whose i^{th} component is $X_p(Y^i)$, for Y^i the i^{th} component of Y . This shows that

$$\Pi(p)(X_p, Y_p) = \langle v(p), X_p(Y) \rangle.$$

If a vector field X is picked similarly, then we have a corresponding equation, and it follows that

$$\Pi(p)(X_p, Y_p) - \Pi(p)(Y_p, X_p) = \langle v(p), X_p(Y) - Y_p(X) \rangle.$$

It is easy to check that in \mathbb{R}^n we have $X_p(Y) - Y_p(X) = [X, Y](p)$, so that

$$\Pi(p)(X_p, Y_p) - \Pi(p)(Y_p, X_p) = \langle v(p), [X, Y](p) \rangle.$$

But the right side is 0, since $[X, Y](p) \in M_p$. ♦

The symmetry of Π states an important property of the map $dv: M_p \rightarrow M_p$; this map is self-adjoint,

$$\langle dv(X), Y \rangle = \langle X, dv(Y) \rangle, \quad X, Y \in M_p.$$

Recall that if V is a vector space with an inner product $\langle \cdot, \cdot \rangle$, then a linear transformation $T: V \rightarrow V$ is called **self-adjoint** (with respect to $\langle \cdot, \cdot \rangle$) if

$$\langle Tv, w \rangle = \langle v, Tw \rangle \quad \text{for all } v, w \in V.$$

This is equivalent to saying that the matrix of T is symmetric with respect to any *orthonormal* basis. It is an elementary fact that eigenvectors v_1 and v_2 of T with distinct eigenvalues must be orthogonal, for if $Tv_i = \lambda_i v_i$ with $\lambda_1 \neq \lambda_2$, then

$$\lambda_1 \langle v_1, v_2 \rangle = \langle Tv_1, v_2 \rangle = \langle v_1, Tv_2 \rangle = \lambda_2 \langle v_1, v_2 \rangle.$$

The main theorem about self-adjoint transformations (the “spectral theorem”) states that a self-adjoint $T: V \rightarrow V$ has a basis of eigenvectors v_1, \dots, v_n . We have seen that eigenvectors with distinct eigenvalues are orthogonal. If two or more eigenvectors have the same eigenvalue, then all vectors in the subspace they span are eigenvectors, so we can select an orthogonal collection spanning the subspace. If we also choose our eigenvectors to be of unit length, we thus obtain an orthonormal basis of eigenvalues. Applying this to $dv: M_p \rightarrow M_p$ we see that there is an orthonormal basis X_1, X_2 of M_p with

$$dv(X_i) = \lambda_i X_i.$$

This fact is what is behind

2. THEOREM (EULER). The curvatures κ_X have a minimum k_1 in one direction and a maximum k_2 in a perpendicular direction. For a direction X making an angle of θ with the first direction we have

$$\kappa_X = k_1 \cos^2 \theta + k_2 \sin^2 \theta.$$

PROOF. Let X_1 and X_2 be unit eigenvectors of Π . Then by Proposition 0

$$\begin{aligned} \kappa_{X_i} &= \Pi(X_i, X_i) = -\langle dv(X_i), X_i \rangle \\ &= -\langle \lambda_i X_i, X_i \rangle \\ &= -\lambda_i. \end{aligned}$$

If we express any other unit vector $X \in M_p$ as

$$X = (\cos \theta)X_1 + (\sin \theta)X_2,$$

then

$$\begin{aligned}\kappa_X &= \Pi(X, X) = -\langle dv(X), X \rangle \\ &= -\langle \lambda_1(\cos \theta)X_1 + \lambda_2(\sin \theta)X_2, (\cos \theta)X_1 + (\sin \theta)X_2 \rangle \\ &= -\lambda_1 \cos^2 \theta - \lambda_2 \sin^2 \theta.\end{aligned}$$

As before, this completes the proof.* ♦

To connect the curvatures k_1, k_2 with K , we first note that K can be expressed very succinctly in terms of dv .

3. PROPOSITION. The Gaussian curvature $K(p)$ at $p \in M$ is

$$K(p) = \text{determinant of } dv: M_p \rightarrow M_p.$$

PROOF. If $Y_1, Y_2 \in M_p$ are linearly independent, then

$$\begin{aligned}K(p) &= \frac{v^*(\sigma')(p)(Y_1, Y_2)}{dV(p)(Y_1, Y_2)} \\ &= \frac{\sigma'(v(p))(v_*Y_1, v_*Y_2)}{dV(p)(Y_1, Y_2)}.\end{aligned}$$

Using the fact that $S^2_{v(p)}$ is parallel to M_p , and remembering that we are considering dv as a map into M_p , this can be written simply

$$K(p) = \frac{dV(p)(dv(Y_1), dv(Y_2))}{dV(p)(Y_1, Y_2)}.$$

Since dV is a 2-form, this ratio is indeed just $\det dv$. ♦

* Notice that we have reduced Euler's Theorem to a fact about the eigenvalues $\lambda_1 \leq \lambda_2$ of a self-adjoint transformation $T: V \rightarrow V$ on a 2-dimensional vector space:

$$\lambda_1 = \min_{\|v\|=1} \langle Tv, v \rangle, \quad \lambda_2 = \max_{\|v\|=1} \langle Tv, v \rangle.$$

For higher dimensions there is a minimax definition of the various eigenvalues. See Courant, *Über die Abhängigkeit ...*, Nachrichten, Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math. Phys. Klasse 1919, pp. 255–264.

4. COROLLARY. Let $Y_1, Y_2 \in M_p$ be orthonormal. Then

$$K(p) = \det(\Pi(Y_i, Y_j)) = \det \begin{pmatrix} \Pi(Y_1, Y_1) & \Pi(Y_1, Y_2) \\ \Pi(Y_2, Y_1) & \Pi(Y_2, Y_2) \end{pmatrix}.$$

PROOF. Since $\Pi(Y_i, Y_j) = \langle -dv(Y_i), Y_j \rangle$, the matrix of dv with respect to Y_1, Y_2 is

$$\begin{pmatrix} -\Pi(Y_1, Y_1) & -\Pi(Y_2, Y_1) \\ -\Pi(Y_1, Y_2) & -\Pi(Y_2, Y_2) \end{pmatrix},$$

which has the same determinant as the matrix $(\Pi(Y_i, Y_j))$. ♦

5. COROLLARY. Let $Y_1, Y_2 \in M_p$ be linearly independent. Then

$$K(p) = \frac{\det(\Pi(Y_i, Y_j))}{\det(\langle Y_i, Y_j \rangle)} = \frac{\det(\Pi(Y_i, Y_j))}{\det(I(Y_i, Y_j))}.$$

PROOF. Corollary 4 proves the formula for orthonormal X_1, X_2 . If $Y_i = \sum_j a_{ij} X_j$, then replacing X_1, X_2 by Y_1, Y_2 multiplies both numerator and denominator by $\det(a_{ij})$. ♦

6. COROLLARY. For any $p \in M$ we have

$$K(p) = k_1 \cdot k_2.$$

PROOF. Apply Corollary 4 when $Y_1, Y_2 \in M_p$ are the orthonormal basis of eigenvectors of dv , with eigenvalues $-k_1, -k_2$. Then

$$K(p) = \det \begin{pmatrix} -k_1 & 0 \\ 0 & -k_2 \end{pmatrix} = k_1 \cdot k_2. \quad \spadesuit$$

Corollary 5 allows us to develop an explicit formula for K , which involves some standard symbolism to be introduced first. If $\chi = (x, y)$ is a coordinate system on M , then we write the first fundamental form I as

$$I = E dx \otimes dx + F dx \otimes dy + F dy \otimes dx + G dy \otimes dy.$$

If $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is the inverse of χ , so that

$$\begin{aligned} \frac{\partial}{\partial x} &= \frac{\partial f}{\partial s} \\ \frac{\partial}{\partial y} &= \frac{\partial f}{\partial t}, \end{aligned}$$

then

$$\begin{aligned}
 E &= \left\langle \frac{\partial f}{\partial s}, \frac{\partial f}{\partial s} \right\rangle = \sum_{i=1}^3 \left(\frac{\partial f^i}{\partial s} \right)^2 \\
 F &= \sum_{i=1}^3 \frac{\partial f^i}{\partial s} \frac{\partial f^i}{\partial t} \\
 G &= \sum_{i=1}^3 \left(\frac{\partial f^i}{\partial t} \right)^2.
 \end{aligned}
 \tag{*}$$

[Notice that the left sides of these equations really mean $E(f(s, t))$, etc. (the functions E, F, G themselves are defined on M). This is just the form we want. For example, to compute $\partial E / \partial x$ at $q = f(\bar{s}, \bar{t})$ we have

$$\begin{aligned}
 \frac{\partial E}{\partial x}(q) &= D_1(E \circ \chi^{-1})(\chi(q)) \\
 &= D_1(E \circ f)(\chi(q)) \\
 &= \frac{\partial E}{\partial s}(\bar{s}, \bar{t});
 \end{aligned}$$

the E on the last line denotes the same function on \mathbb{R}^2 which appears in the equations (*).

The symbols E, F, G were introduced by Gauss himself (at the beginning of section 11 of his paper), and they have remained standard ever since. There are also standard symbols for the second fundamental form:

$$\Pi = l \, dx \otimes dx + m \, dx \otimes dy + m \, dy \otimes dx + n \, dy \otimes dy.$$

To obtain formulas for l, m, n , we look at the Weingarten equations in the first proof of Theorem 1, and note that

$$\begin{aligned}
 l &= \Pi \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial s} \right) = \left\langle N, \frac{\partial^2 f}{\partial s^2} \right\rangle = \sum_{i=1}^3 v^i(f(s, t)) \frac{\partial^2 f^i}{\partial s^2} \\
 m &= \sum_{i=1}^3 v^i(f(s, t)) \frac{\partial^2 f^i}{\partial s \partial t} \\
 n &= \sum_{i=1}^3 v^i(f(s, t)) \frac{\partial^2 f^i}{\partial t^2};
 \end{aligned}
 \tag{**}$$

formulas for $v^i(f(s, t))$ have already been given on pages 113–114. Using Corollary 5, we now have the classical formula

$$K(p) = \frac{ln - m^2}{EG - F^2}(p).$$

The symbols l, m, n do not appear in Gauss, who instead uses the symbols D, D', D'' for certain quantities proportional to them. These symbols are introduced on page 87; it is easy to see that Gauss' formula for K on this page is equivalent to the one we have just derived, although Gauss obtained it in a different way, by beginning with the formula which we derived on page 119.

Our next theorem probably requires an apology in advance. The result looks amazingly unappetizing; it's hard to see why anyone would want it even if he had it, and the proof is merely an involved calculation. Nevertheless, we will justify its existence soon after proving it. The calculation appearing in the proof should be a lot easier to follow than Gauss'.

7. THEOREM. Let (x, y) be a coordinate system on a neighborhood of $p \in M \subset \mathbb{R}^3$, and let

$$I = i^*(\langle \cdot, \cdot \rangle) = E dx \otimes dx + F dx \otimes dy + F dy \otimes dx + G dy \otimes dy.$$

Then

$$\begin{aligned} 4(EG - F^2)^2 K = & E \left(\frac{\partial E}{\partial y} \frac{\partial G}{\partial y} - 2 \frac{\partial F}{\partial x} \frac{\partial G}{\partial y} + \left(\frac{\partial G}{\partial x} \right)^2 \right) \\ & + F \left(\frac{\partial E}{\partial x} \frac{\partial G}{\partial y} - \frac{\partial E}{\partial y} \frac{\partial G}{\partial x} - 2 \frac{\partial E}{\partial y} \frac{\partial F}{\partial y} + 4 \frac{\partial F}{\partial x} \frac{\partial F}{\partial y} - 2 \frac{\partial F}{\partial x} \frac{\partial G}{\partial x} \right) \\ & + G \left(\frac{\partial E}{\partial x} \frac{\partial G}{\partial x} - 2 \frac{\partial E}{\partial x} \frac{\partial F}{\partial y} + \left(\frac{\partial E}{\partial y} \right)^2 \right) \\ & - 2(EG - F^2) \left(\frac{\partial^2 E}{\partial y^2} - 2 \frac{\partial^2 F}{\partial x \partial y} + \frac{\partial^2 G}{\partial x^2} \right). \end{aligned}$$

PROOF. Consider the inverse $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ of the coordinate system (x, y) . To save space we will denote

$$\frac{\partial^2 f}{\partial s \partial t} = D_{12} f \quad \text{simply by } f_{12},$$

and similarly for other partial derivatives. We have

$$K = \frac{ln - m^2}{EG - F^2},$$

where, by (**),

$$\begin{aligned} l &= \langle f_{11}, N \rangle = \left\langle f_{11}, \frac{f_1 \times f_2}{\sqrt{EG - F^2}} \right\rangle \\ m &= \langle f_{12}, N \rangle = \left\langle f_{12}, \frac{f_1 \times f_2}{\sqrt{EG - F^2}} \right\rangle \\ n &= \langle f_{22}, N \rangle = \left\langle f_{22}, \frac{f_1 \times f_2}{\sqrt{EG - F^2}} \right\rangle. \end{aligned}$$

Thus

$$\begin{aligned} K(EG - F^2)^2 &= \langle f_{11}, f_1 \times f_2 \rangle \cdot \langle f_{22}, f_1 \times f_2 \rangle - \langle f_{12}, f_1 \times f_2 \rangle^2 \\ &= \det \begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix} \cdot \det \begin{pmatrix} f_{22} \\ f_1 \\ f_2 \end{pmatrix} - \det \begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix} \cdot \det \begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix}. \end{aligned}$$

In this equation each f_i and f_{ij} is considered as a row of the matrix. If we use f_i^t and f_{ij}^t to denote the columns with the same entries, then we also have

$$\begin{aligned} K(EG - F^2)^2 &= \det \begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix} \cdot \det(f_{22}^t, f_1^t, f_2^t) - \det \begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix} \cdot \det(f_{12}^t, f_1^t, f_2^t) \\ &= \det \left[\begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix} \cdot (f_{22}^t, f_1^t, f_2^t) \right] \\ &\quad - \det \left[\begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix} \cdot (f_{12}^t, f_1^t, f_2^t) \right] \\ &= \det \begin{pmatrix} \langle f_{11}, f_{22} \rangle & \langle f_{11}, f_1 \rangle & \langle f_{11}, f_2 \rangle \\ \langle f_1, f_{22} \rangle & E & F \\ \langle f_2, f_{22} \rangle & F & G \end{pmatrix} \\ &\quad - \det \begin{pmatrix} \langle f_{12}, f_{12} \rangle & \langle f_{12}, f_1 \rangle & \langle f_{12}, f_2 \rangle \\ \langle f_{12}, f_1 \rangle & E & F \\ \langle f_{12}, f_2 \rangle & F & G \end{pmatrix} \end{aligned}$$

$$= \det \begin{pmatrix} \langle f_{11}, f_{22} \rangle - \langle f_{12}, f_{12} \rangle & \langle f_{11}, f_1 \rangle & \langle f_{11}, f_2 \rangle \\ \langle f_1, f_{22} \rangle & E & F \\ \langle f_2, f_{22} \rangle & F & G \end{pmatrix} \\ - \det \begin{pmatrix} 0 & \langle f_{12}, f_1 \rangle & \langle f_{12}, f_2 \rangle \\ \langle f_{12}, f_1 \rangle & E & F \\ \langle f_{12}, f_2 \rangle & F & G \end{pmatrix}.$$

But from the definitions of E, F, G in (*) we have

$$\langle f_{11}, f_1 \rangle = \frac{1}{2} E_1 \quad (E_1 = \partial E / \partial s)$$

$$\langle f_{12}, f_1 \rangle = \frac{1}{2} E_2$$

$$\langle f_{22}, f_2 \rangle = \frac{1}{2} G_2$$

$$\langle f_{12}, f_2 \rangle = \frac{1}{2} G_1$$

$$\langle f_{11}, f_2 \rangle = F_1 - \frac{1}{2} E_2$$

$$\langle f_{22}, f_1 \rangle = F_2 - \frac{1}{2} G_1.$$

Moreover, from the fourth and fifth equations we obtain

$$\frac{1}{2} G_{11} = \frac{\partial}{\partial s} \langle f_{12}, f_2 \rangle = \langle f_{121}, f_2 \rangle + \langle f_{12}, f_{21} \rangle$$

$$F_{12} - \frac{1}{2} E_{22} = \frac{\partial}{\partial t} \langle f_{11}, f_2 \rangle = \langle f_{112}, f_2 \rangle + \langle f_{11}, f_{22} \rangle.$$

Subtracting the first of these from the second, we then obtain

$$\langle f_{11}, f_{22} \rangle - \langle f_{12}, f_{12} \rangle = -\frac{1}{2} G_{11} + F_{12} - \frac{1}{2} E_{22}.$$

So we have, finally,

$$K(EG - F^2)^2 = \det \begin{pmatrix} -\frac{1}{2} G_{11} + F_{12} - \frac{1}{2} E_{22} & \frac{1}{2} E_1 & F_1 - \frac{1}{2} E_2 \\ F_2 - \frac{1}{2} G_1 & E & F \\ \frac{1}{2} G_2 & F & G \end{pmatrix} \\ - \det \begin{pmatrix} 0 & \frac{1}{2} E_2 & \frac{1}{2} G_1 \\ \frac{1}{2} E_2 & E & F \\ \frac{1}{2} G_1 & F & G \end{pmatrix},$$

which gives the formula in the statement of the theorem! ❖

From Theorem 7 we can deduce an immediate Corollary, which appears on page 93 of the translation:

“Thus the formula of the preceding article leads of itself to the remarkable

THEOREM. If a curved surface is developed upon any other surface whatever, the measure of curvature in each point remains unchanged.”

The Latin word for ‘remarkable’ has become part of the traditional name for this result:

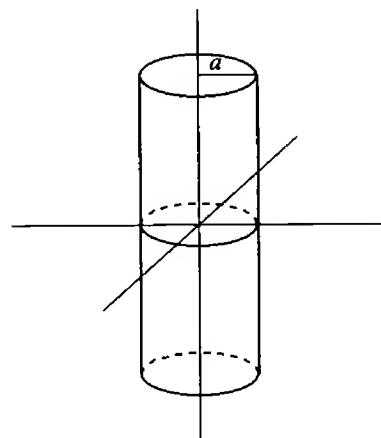
8. COROLLARY (THEOREMA EGREGIUM). If $f, g: M \rightarrow \mathbb{R}^3$ are two imbeddings (or even immersions) such that $f^*\langle \cdot, \cdot \rangle = g^*\langle \cdot, \cdot \rangle$, then the Gaussian curvature of $f(M) \subset \mathbb{R}^3$ at $f(p)$ equals the Gaussian curvature of $g(M)$ at $g(p)$.

The Theorema Egregium justifies the close attention which we have given to the Gaussian curvature K of a surface. Although defined in terms of the imbedding of the surface in \mathbb{R}^3 , it turns out to depend only on the Riemannian metric induced by that imbedding. This shows why the Gaussian curvature of a cylinder must be 0—there is a (local) isometry from the plane to the cylinder: If the cylinder $Z \subset \mathbb{R}^3$ is

$$Z = \{(x, y, z) : x^2 + y^2 = a^2\},$$

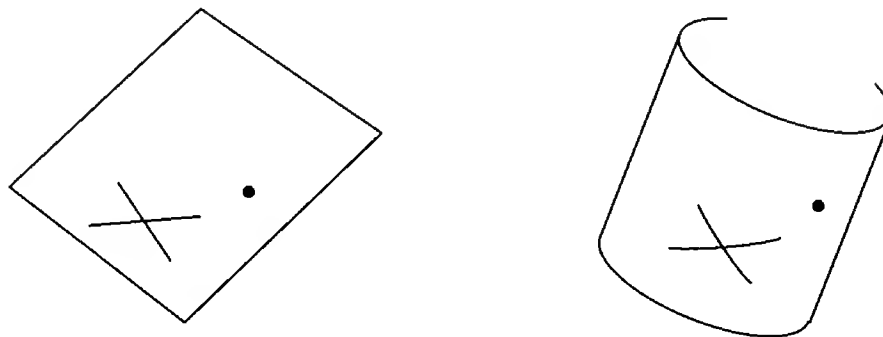
then a local isometry $f: \mathbb{R}^2 \rightarrow Z$ is given by

$$f(s, t) = \left(a \cos \frac{s}{a}, a \sin \frac{s}{a}, t \right).$$

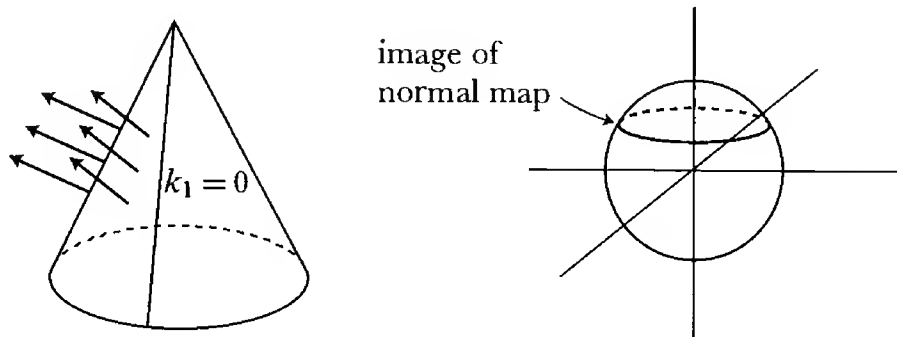


It is easily checked that f_* is an isometry at any point, but the result should be clear without any calculations whatsoever. To prove that Z is locally isometric to \mathbb{R}^2 , just take a piece of paper, and roll it up into a cylinder. The map which takes a point on the flat piece of paper into the corresponding point on the rolled up piece of paper is an isometry. The isometric properties of this map

are expressed by the everyday experience that paper cannot be stretched, but merely bent.

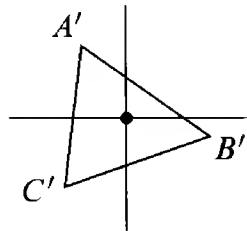
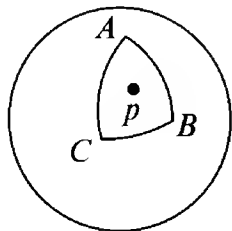


The Theorema Egregium is often expressed by saying that Gaussian curvature is a “bending invariant”. Anyone who has ever made a paper dunce hat knows (though he may not know that he knows) that the cone is also locally isometric to the plane, and hence has Gaussian curvature 0.



Map makers, and anyone who has had to wrap a spherical object, know that a piece of paper cannot be bent onto even a small portion of a sphere. A mathematical proof follows immediately from the Theorema Egregium, for a sphere has non-zero Gaussian curvature. The situation for surfaces is thus completely different from that for curves. All 1-dimensional Riemannian manifolds are locally isometric to \mathbb{R}^1 , for if we choose an immersed curve c in the manifold, then the arclength function of c is an isometry into \mathbb{R}^1 . So there are no interesting bending invariants of a curve; all the interesting characteristics of a curve are invariants under the group of Euclidean motions. The Gaussian curvature is, of course, an invariant under the group of Euclidean motions, but it is also invariant under the much larger (but still important) group of maps which are merely defined on the surface, and are isometries there. (As a contrast, the **mean curvature** $\frac{1}{2}(k_1 + k_2)$ is invariant under the group of Euclidean motions, but it is *not* a bending invariant; for example, the plane has mean curvature 0, while a cylinder of radius a has mean curvature $a/2$.)

In the previous paragraph we appealed to the Theorema Egregium to prove that a sphere is not locally isometric to the plane. But it is also possible to give a much more elementary proof of this fact, that will convince some one who knows a little geometry. If there were an isometry from a neighborhood of $p \in S^2$ into the plane, then a small triangle ABC around p , with portions



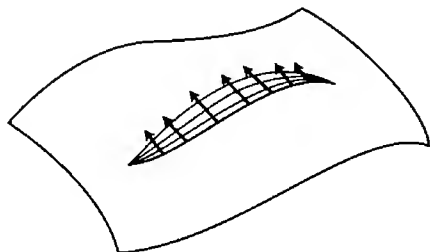
of great circles as sides, would have to be mapped into an ordinary triangle $A'B'C'$ on the plane, since great circles are geodesics on the sphere. The angles at A, B, C would also have to equal the angles A', B', C' . This is impossible, since $\angle A + \angle B + \angle C > \pi$, while $\angle A' + \angle B' + \angle C' = \pi$. This phenomenon turns out to have a generalization to arbitrary surfaces, a result that Gauss felt “ought to be counted among the most elegant in the theory of curved surfaces”. In deriving this result we will essentially follow Gauss, but we will use some of our previous results about geodesics, and suppress some of the additional formulas which Gauss obtains along the way, so that the argument may appear somewhat simpler.

In expounding the theory of geodesics, Gauss derives two conditions, of entirely different natures, for a geodesic on a surface. Although the first of these conditions is not necessary for our final goal, it is an interesting exercise in the calculus of variations, as well as an interesting result in its own right. We consider a curve $c: [a, b] \rightarrow M$ and a variation α keeping endpoints fixed. Looking at the energy function (Gauss looks at length instead), we have

$$\begin{aligned} \frac{dE(\bar{\alpha}(u))}{du} \Big|_{u=0} &= \frac{d}{du} \Big|_{u=0} \frac{1}{2} \int_a^b \left\langle \frac{\partial \alpha}{\partial t}(u, t), \frac{\partial \alpha}{\partial t}(u, t) \right\rangle dt \\ &= \int_a^b \left\langle \gamma'(t), \frac{\partial^2 \alpha}{\partial u \partial t}(0, t) \right\rangle dt \\ &= - \int_a^b \left\langle \gamma''(t), \frac{\partial \alpha}{\partial u}(0, t) \right\rangle dt, \quad \text{using integration by parts.} \end{aligned}$$

This result holds for *any* variation α keeping endpoints fixed. The final integral

must therefore be zero whenever $\partial\alpha/\partial u(0,t) \in M_{\gamma(t)}$ for all t , since there is then a variation α through curves on M with these values of $\partial\alpha/\partial u(0,t)$. In other



words, γ is a geodesic on M if and only if

$$\int_a^b \langle \gamma''(t), \eta(t) \rangle dt = 0 \quad \text{for every } \eta \text{ satisfying } \langle \eta(t), v(\gamma(t)) \rangle = 0.$$

In particular, we can choose

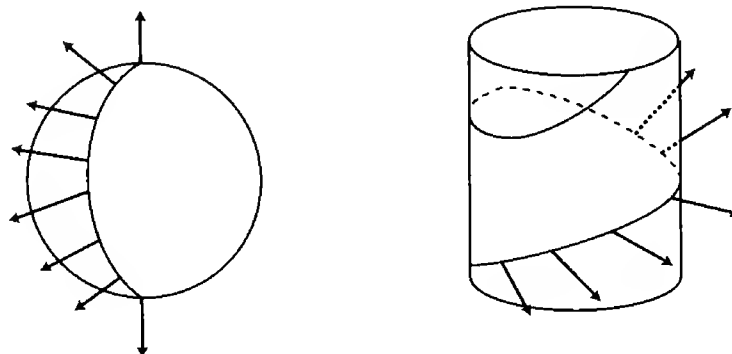
$$\eta(t) = \phi(t)[\gamma''(t) - \langle \gamma''(t), v(\gamma(t)) \rangle v(\gamma(t))]$$

where ϕ is a C^∞ function on $[a, b]$ with $\phi > 0$ on (a, b) and $\phi(a) = \phi(b) = 0$. We then obtain

$$0 = \int_a^b \phi(t) [\langle \gamma''(t), \gamma''(t) \rangle - \langle \gamma''(t), v(\gamma(t)) \rangle^2] dt.$$

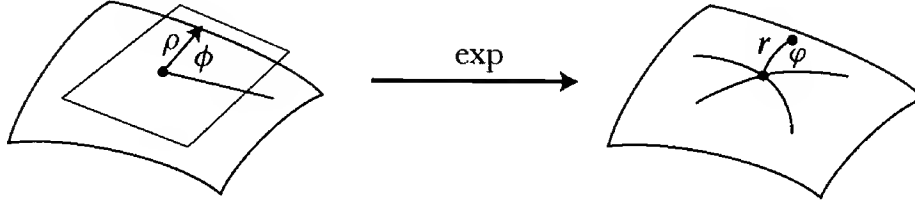
Since $v(\gamma(t))$ has length 1, the Schwarz inequality (Theorem I.9-1(2)) shows that the term in brackets is always ≤ 0 . Since $\phi(t) > 0$ on (a, b) , the term in brackets must actually be 0 everywhere. This implies that $\gamma''(t)$ and $v(\gamma(t))$ are everywhere *linearly dependent*. In other words,

The curve γ on M is a geodesic if and only if
 γ'' is always perpendicular to M .



Notice that this condition makes sense only for a surface in \mathbb{R}^3 ; the vector γ'' has no meaning for an abstract Riemannian manifold.

We now pass to the equations for a geodesic on any 2-dimensional Riemannian manifold. However, we will work with a special coordinate system. Consider a neighborhood of $p \in M$ which is $\exp(U)$, where $U \subset M_p$ is a neighborhood of 0 on which \exp is one-one. Identify M_p with \mathbb{R}^2 by choosing an orthonormal basis for M_p . Introducing polar coordinates (ρ, ϕ) on M_p (minus some ray), we obtain a coordinate system $(r, \varphi) = (\rho, \phi) \circ \exp^{-1}$ on $\exp(U)$ (minus some geodesic ray), where ρ is now chosen so that $\rho = 1$ for vectors in M_p with norm 1.



This implies that $\|\partial/\partial r\| = 1$. We also know that $\langle \partial/\partial r, \partial/\partial \varphi \rangle = 0$, by Gauss' Lemma. So we have

$$\langle \cdot, \cdot \rangle = dr \otimes dr + G d\varphi \otimes d\varphi$$

for some function G . The function G is just $G(q) = \langle \partial/\partial \varphi|_q, \partial/\partial \varphi|_q \rangle$. Clearly, G can be considered as defined on $\exp(U) - \{p\}$, even though any particular coordinate system (r, φ) can be defined only on $\exp(U)$ minus some geodesic ray. In terms of the g_{ij} notation we have

$$\begin{aligned} g_{11} &= 1, & g^{11} &= 1 \\ g_{12} &= g_{21} = 0, & g^{12} &= g^{21} = 0 \\ g_{22} &= G, & g^{22} &= \frac{1}{G}. \end{aligned}$$

An easy calculation from the definitions (pp. I.326 and I.328) then gives

$$[12, 2] = [22, 2] = \frac{1}{2} \frac{\partial G}{\partial r}$$

$$[22, 1] = -\frac{1}{2} \frac{\partial G}{\partial r}$$

$$\text{all other } [ij, k] = 0;$$

$$\Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{2G} \frac{\partial G}{\partial r}$$

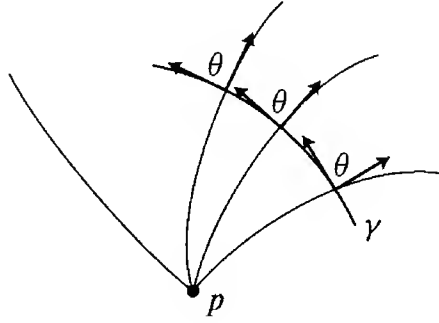
$$\Gamma_{22}^1 = -\frac{1}{2} \frac{\partial G}{\partial r}$$

$$\text{all other } \Gamma_{ij}^k = 0.$$

The equations for a geodesic (pg. I.329) thus give

$$(1) \quad \frac{d^2 \gamma^1}{ds^2} = \frac{1}{2} \frac{\partial G}{\partial r}(\gamma(s)) \left(\frac{d\gamma^2}{ds} \right)^2.$$

Now suppose γ is some curve parameterized by arclength, and let $\theta(s)$ be the angle between $\gamma'(s)$ and $\partial/\partial r|_{\gamma(s)}$. Clearly



$$(2) \quad \begin{aligned} \cos \theta(s) &= \left\langle \gamma'(s), \frac{\partial}{\partial r} \Big|_{\gamma(s)} \right\rangle \\ &= \left\langle \frac{d\gamma^1}{ds} \frac{\partial}{\partial r} \Big|_{\gamma(s)} + \frac{d\gamma^2}{ds} \frac{\partial}{\partial \varphi} \Big|_{\gamma(s)}, \frac{\partial}{\partial r} \Big|_{\gamma(s)} \right\rangle \\ &= \frac{d\gamma^1}{ds}. \end{aligned}$$

For a geodesic γ we obtain from (1) and (2),

$$(3) \quad \begin{aligned} \frac{1}{2} \frac{\partial G}{\partial r}(\gamma(s)) \left(\frac{d\gamma^2}{ds} \right)^2 &= \frac{d \cos \theta(s)}{ds} \\ &= -\sin \theta(s) \frac{d\theta(s)}{ds}. \end{aligned}$$

Finally, note that the area of the parallelogram spanned by the unit vectors $\gamma'(s)$ and $\partial/\partial r|_{\gamma(s)}$ equals

$$(4) \quad \begin{aligned} \sin \theta(s) \quad \text{and also equals} \quad dV \left(\frac{\partial}{\partial r} \Big|_{\gamma(s)}, \gamma'(s) \right) \\ &= \sqrt{G}(\gamma(s)) dr \wedge d\varphi \left(\frac{\partial}{\partial r} \Big|_{\gamma(s)}, \frac{d\gamma^1}{ds} \frac{\partial}{\partial r} \Big|_{\gamma(s)} + \frac{d\gamma^2}{ds} \frac{\partial}{\partial \varphi} \Big|_{\gamma(s)} \right) \\ &= \sqrt{G}(\gamma(s)) \frac{d\gamma^2}{ds}. \end{aligned}$$

So we obtain

$$\frac{1}{2} \frac{\partial G}{\partial r}(\gamma(s)) \left(\frac{d\gamma^2}{ds} \right)^2 = -\sqrt{G}(\gamma(s)) \frac{d\gamma^2}{ds} \frac{d\theta}{ds},$$

$$\frac{d\theta}{ds} = -\frac{1}{2} \frac{\frac{\partial G}{\partial r}}{\sqrt{G}}(\gamma(s)) \frac{d\gamma^2}{ds},$$

and thus

$$(*) \quad \frac{d\theta}{ds} = \frac{\partial \sqrt{G}}{\partial r}(\gamma(s)) \frac{d\gamma^2}{ds}$$

for any geodesic γ . This is the equation Gauss finally obtains, on page 107 of the translation.

Gauss also obtains the expression for K in this special coordinate system. Theorem 7 now takes the much simpler form

$$4G^2 K = \left(\frac{\partial G}{\partial r} \right)^2 - 2G \frac{\partial^2 G}{\partial r^2},$$

which gives

$$(**) \quad K = -\frac{1}{\sqrt{G}} \frac{\partial^2 \sqrt{G}}{\partial r^2}.$$

It will now be necessary to obtain some further information about the function \sqrt{G} , for which Gauss gives very brief arguments. We will find it convenient to express G “as a function of ρ and ϕ ”; that is, we consider

$$g = G \circ \exp \circ P^{-1},$$

where $P: M_p\text{-ray} \rightarrow (0, \delta] \times (0, 2\pi)$ is $P = (\rho, \phi)$, so that $g(\rho_0, \phi_0)$ is $G(\exp v)$, where $v \in M_p$ has polar coordinates (ρ_0, ϕ_0) . At times it will also be convenient to use (ρ, ϕ) to stand for a point in M_p , as well as standing for the coordinate functions themselves.

The function g can be considered as defined on $(0, \delta] \times [0, 2\pi]$ (with the same values at $(\rho, 0)$ as at $(\rho, 2\pi)$). What we want to examine is the behavior

of g near $(0, \phi)$. We do this by comparing distances on M_p with those on M . On the vector space M_p we have an inner product $\langle \cdot, \cdot \rangle_p$; since the tangent space $(M_p)_v$ of M_p at $v \in M_p$ can be identified with M_p , we can use $\langle \cdot, \cdot \rangle_p$ to obtain a Riemannian metric on M_p . To keep things straight, we will use v, w to denote elements of M_p , and X_v, Y_v to denote tangent vectors in the tangent space $(M_p)_v$. For each $X_v \in (M_p)_v$ we thus have a certain norm, which we will denote by $\|X_v\|$. Now recall that $\exp_*: (M_p)_0 \rightarrow M_p$ is the identity map (when we identify $(M_p)_0$ with M_p), so that we certainly have

$$\|X_0\| = \|\exp_*(X_0)\|_p, \quad X_0 \in (M_p)_0.$$

If $\varepsilon > 0$, it follows that for $v \in M_p$ sufficiently close to 0 and $X_v \in (M_p)_v$ of unit norm $\|X_v\| = 1$ we have

$$\left| \|X_v\| - \|\exp_*(X_v)\|_{\exp(v)} \right| < \varepsilon,$$

so that for any $Y_v \in (M_p)_v$ we have

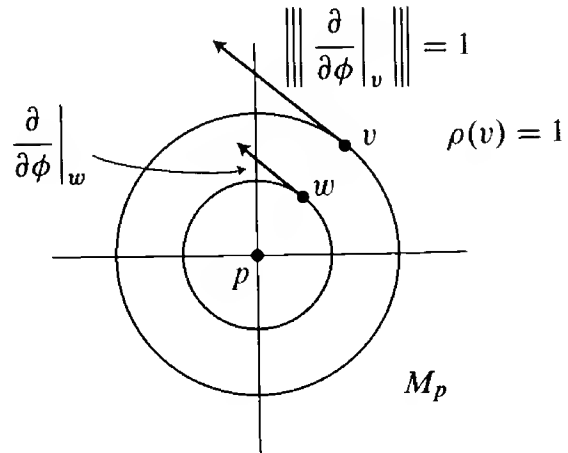
$$\left| \|Y_v\| - \|\exp_*(Y_v)\|_{\exp(v)} \right| < \varepsilon \cdot \|Y_v\|.$$

Noting that $\exp_*(\partial/\partial\phi|_v) = \partial/\partial\phi|_{\exp(v)}$, we have

$$(1) \quad \left| \left\| \frac{\partial}{\partial\phi} \right\|_v - \left\| \frac{\partial}{\partial\phi} \right\|_{\exp(v)} \right| < \varepsilon \cdot \left\| \frac{\partial}{\partial\phi} \right\|_v,$$

for all sufficiently small v , while clearly

$$\left\| \frac{\partial}{\partial\phi} \right\|_v = \|v\|_p = \rho(v).$$



Dividing all terms of (1) by ρ yields

$$\left| 1 - \frac{\sqrt{g}(\rho, \phi)}{\rho} \right| < \varepsilon,$$

for all sufficiently small ρ . Since this is true for all $\varepsilon > 0$, we have thus shown that

$$(2) \quad \sqrt{g}(\rho, \phi) = \rho + o(\rho),$$

where $o(\rho)$ denotes a function on $(0, \delta] \times [0, 2\pi]$ such that

$$\lim_{\rho \rightarrow 0} \frac{o(\rho)}{\rho} = 0 \quad (\text{uniformly in } \phi).$$

Clearly \sqrt{g} remains continuous on $[0, \delta] \times [0, 2\pi]$ if we define

$$(***) \quad \sqrt{g}(0, \phi) = 0.$$

Notice, moreover, that equation (2) now immediately implies that

$$(***) \quad \frac{\partial \sqrt{g}}{\partial \rho}(0, \phi) = 1$$

(where $\partial \sqrt{g} / \partial \rho(0, \phi)$ really denotes a right hand derivative). However, we want to know that $\partial \sqrt{g} / \partial \rho$ is actually continuous on $[0, \delta] \times [0, 2\pi]$; the argument for this will require another step.

From equation (**) we have, on $(0, \delta] \times [0, 2\pi]$,

$$\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) = -\sqrt{g}(\rho, \phi) \cdot K(\exp(\rho, \phi)) \quad \begin{array}{l} \text{[where } \exp(\rho, \phi) \text{ really means} \\ \exp(v), \text{ where } v \in M_p \text{ has} \\ \text{polar coordinates } (\rho, \phi)\text{]}, \end{array}$$

which shows that $\partial^2 \sqrt{g} / \partial \rho^2(\rho, \phi) \rightarrow 0$ as $\rho \rightarrow 0$. It follows, in particular, that $\partial^2 \sqrt{g} / \partial \rho^2$ is bounded. For $\rho > 0$ we have

$$\frac{\partial \sqrt{g}}{\partial \rho}(\rho, \phi) = \frac{\partial \sqrt{g}}{\partial \rho}(\delta, \phi) - \int_{\rho}^{\delta} \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(t, \phi) dt,$$

which immediately implies that

$$\lim_{\rho \rightarrow 0} \frac{\partial \sqrt{g}}{\partial \rho}(\rho, \phi) \text{ exists.}$$

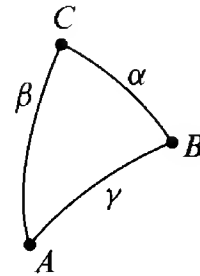
By a standard theorem of calculus (see e.g., Spivak, *Calculus*, 3rd ed., pg. 200), the limit must be $\partial \sqrt{g} / \partial \rho(0, \phi)$, which equals 1, by (****). It follows, in particular, that

$$\begin{aligned} \text{(*****)} \quad \int_0^{\rho_0} -\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) d\rho &= \lim_{\varepsilon \rightarrow 0} \int_{\varepsilon}^{\rho_0} -\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) d\rho \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\partial \sqrt{g}}{\partial \rho}(\varepsilon, \phi) - \frac{\partial \sqrt{g}}{\partial \rho}(\rho_0, \phi) \\ &= 1 - \frac{\partial \sqrt{g}}{\partial \rho}(\rho_0, \phi). \end{aligned}$$

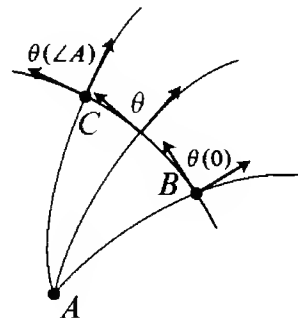
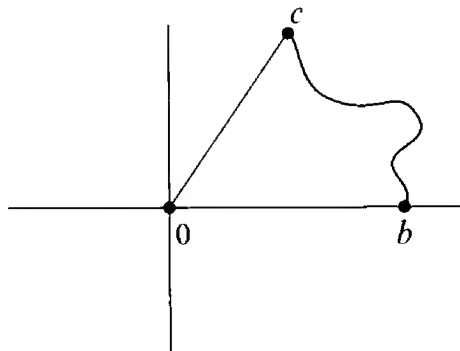
We are ready to prove a theorem.

9. THEOREM. Let A, B, C be three points of $\exp(U)$, where U is a convex neighborhood of $0 \in M_A$, on which \exp is a diffeomorphism, and let α be a geodesic in $\exp(U)$ between B and C . Denote the geodesic segment from A to C by β and the geodesic segment from A to B by γ , and let $\triangle ABC$ be the “geodesic triangle” bounded by α, β, γ . Also let $\angle A$ denote the angle between β and γ , etc. Then

$$\int_{\triangle ABC} K dV = \angle A + \angle B + \angle C - \pi.$$



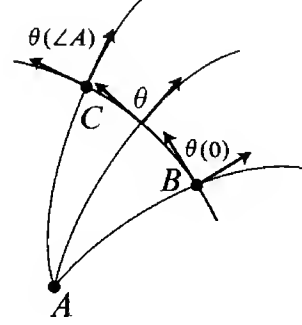
PROOF. Let $b, c \in M_p$ be the vectors with $\exp(b) = B$ and $\exp(c) = C$. Choose the polar coordinates ρ, ϕ on M_p so that $\phi(b) = 0$. Then $\phi(c)$ is just $\angle A$.



It is easy to see that α cannot intersect the same geodesic ray through A twice, so it must be the image under \exp of a curve in M_A which is the graph $\rho = f(\phi)$ of some function in the polar coordinates ρ, ϕ . Let $\theta(\phi)$ be the angle between $\partial/\partial r|_q$ and $\alpha'(q)$, where q is \exp of the vector with polar coordinates $(f(\phi), \phi)$.

We clearly have

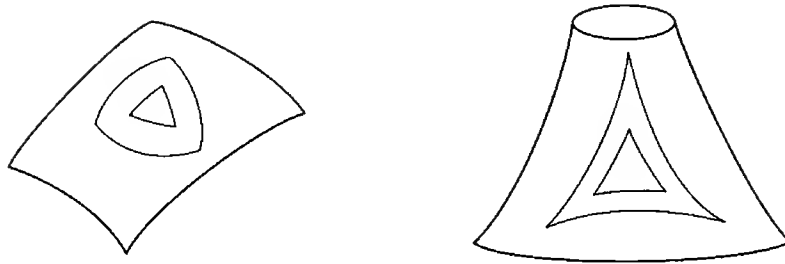
$$\theta(0) = \pi - \angle B, \quad \theta(\angle A) = \angle C.$$



Consequently,

$$\begin{aligned}
 \int_{\Delta ABC} K dV &= \int_{\exp^{-1}(\Delta ABC)} \exp^*(K dV) \\
 &= \int_{\exp^{-1}(\Delta ABC)} -\frac{1}{\sqrt{g}} \frac{\partial^2 \sqrt{g}}{\partial \rho^2} \cdot \sqrt{g} d\rho \wedge d\phi \quad \text{by } (**) \\
 &= \int_0^{\angle A} \left(\int_0^{f(\phi)} -\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) d\rho \right) d\phi \\
 &= \int_0^{\angle A} \left(1 - \frac{\partial \sqrt{g}}{\partial \rho}(f(\phi), \phi) \right) d\phi \quad \text{by } (****) \\
 &= \int_0^{\angle A} \left(1 + \frac{d\theta}{d\phi}(\phi) \right) d\phi \quad \text{by } (*) \\
 &= \angle A + \theta(\angle A) - \theta(0) \\
 &= \angle A + \angle C + \angle B - \pi. \quad \spadesuit
 \end{aligned}$$

According to Theorem 9, on a surface with everywhere positive curvature the sum of the angles of a triangle with geodesic sides is always $> \pi$, while on a surface with everywhere negative curvature the sum of the angle is always $< \pi$. This certainly looks like the case in pictures, and one can even see that the bigger the triangle, the bigger the difference between $\angle A + \angle B + \angle C$ and π .



Notice that in the proof of Theorem 9 we are essentially converting an integral over the region $\triangle ABC$ into an integral over (one of) its sides—this looks suspiciously like Stokes' Theorem. In Volume III, we will indeed be able to present a much nicer proof of Theorem 9, which does not depend on a special coordinate system, and which makes explicit the role of Stokes' Theorem. Moreover, we will be able to derive other important consequences of the same results. However, for the moment, we are more interested in two questions. How did Gauss think of the Theorema Egregium? and What does it really mean?

The paper which we have just examined was based on an earlier paper which Gauss did not publish (also included in the Princeton University Library translation). From the earlier paper it appears that Gauss first proved the result in Theorem 9. Notice that this result gives the Theorema Egregium as a corollary, for it implies that

$$K(p) = \lim_{\triangle ABC \rightarrow p} \frac{\angle A + \angle B + \angle C - \pi}{\text{area } \triangle ABC},$$

and this limit is defined totally in terms of the Riemannian metric on the surface. After realizing this, Gauss probably said to himself: "Since K depends only on the metric, it should be possible to show this by a direct computation; and if any one can do the computation, I certainly can."

To answer the second question—What does the Theorema Egregium really mean?—requires a more serious effort. The original definition of K seemed perfectly satisfactory—it has immediate geometric appeal and is even fairly easy to compute. The only defect of the definition is that the concept being defined turns out to be too good; it turns out to be invariant under isometries, while its definition is not.

One of the dogmas of modern mathematics is that for any object that is invariant in any sort of way, a definition must be found that exhibits this invariance directly (even if such a definition is harder to understand than the original!). The definition of determinants is a good example. An elementary treatment of determinants usually begins by defining the determinant of a *matrix*, either by writing down a messy formula, or in an inductive way that really amounts to the messy formula. It is then shown that $\det(A \cdot B) = \det A \cdot \det B$; from this

it follows that one can define the determinant of a *linear transformation* to be the determinant of its matrix with respect to any basis. This naturally leads one to seek a definition of the determinant of a linear transformation $T: V \rightarrow V$ which does not require a choice of basis. After one has defined $\Omega^k(V)$, it is possible to define $\det T$ to be the constant such that $T^*: \Omega^n(V) \rightarrow \Omega^n(V)$ is $\det T$ times the identity. This definition is indeed independent of a choice of basis, but when one looks a little harder at it, one sees that it isn't all that different from the messy formula defining the determinant of a matrix. Indeed, one proves that $\dim \Omega^n(V) \neq 0$ by writing down an explicit non-zero element of it (in terms of a basis!) which involves permutations in exactly the same way as the original definition of determinants. Finally, if one can tolerate even more complicated constructions, the “exterior algebra” of V can be used to produce a definition of $\det T$ which is completely independent of bases, and does not even mention permutations. In fact, the sign of a permutation σ can then be *defined* as the determinant of the linear transformation $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ with $T(e) = e_{\sigma(i)}$. This approach is expounded in Chevalley's book, *The Construction and Study of Certain Important Algebras*.

The definition of curvature follows a similar course. We first defined the curvature of a surface in a way which depends on an imbedding in \mathbb{R}^3 ; our final goal is a definition of curvature which depends only on the Riemannian metric on the surface. It should be noted that we already have a candidate—the formula in the Theorema Egregium may be used as a definition of curvature! It must still be checked that this definition does not depend on the coordinate system (an uninviting task), but at least the definition involves only a coordinate system on the surface, not an imbedding of the surface in \mathbb{R}^3 . Presumably, few mathematicians would accept this definition as a reasonable one. As we proceed to frame more acceptable definitions of curvature, we will rely, just as in the case of determinants, on more complicated and abstract constructions, so it is important that we follow the historical evolution of the definition of curvature, in order not to lose sight of its geometric significance.

One further comparison with the case of determinants will point out the direction which our investigations will have to take. It seems safe to assert that the modern invariant definitions of determinants would never have come into being if mathematicians had always considered only determinants of 2×2 matrices—our perception of the structure is complicated too much by the simplicity of this special case. Similarly for curvature. The essence of curvature is revealed more fully only when one transcends the limits of intuition and considers manifolds of arbitrary dimensions, conceived as existing in their own right, not as subsets of a Euclidean space.

ADDENDUM
THE FORMULA OF
BERTRAND AND PUISEUX;
DIQUET'S FORMULA

Consider the function \sqrt{g} which was used in the proof of Theorem 9. This function is defined on some $[0, \delta] \times [0, 2\pi]$ and

$$(1) \quad \sqrt{g}(\rho, 0) = \sqrt{g}(\rho, 2\pi) \quad \text{for } \rho \in [0, \delta]$$

$$(2) \quad \sqrt{g}(0, \phi) = 0$$

$$(3) \quad \frac{\partial \sqrt{g}}{\partial \rho}(0, \phi) = \lim_{\rho \rightarrow 0} \frac{\partial \sqrt{g}}{\partial \rho}(\rho, \phi) = 1.$$

We have also noted that on $(0, \delta] \times [0, 2\pi]$ we have

$$(4) \quad \begin{aligned} \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) &= -\sqrt{g}(\rho, \phi) \cdot K(\exp(\rho, \phi)) \\ &= -\sqrt{g}(\rho, \phi) \cdot \bar{K}(\rho, \phi), \quad \text{say.} \end{aligned}$$

This shows (appealing once again to the standard theorem of calculus used before) that

$$(5) \quad \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(0, \phi) = \lim_{\rho \rightarrow 0} \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) = 0.$$

Differentiating equation (4) yields

$$(6) \quad \frac{\partial^3 \sqrt{g}}{\partial \rho^3}(\rho, \phi) = -\frac{\partial \sqrt{g}}{\partial \rho}(\rho, \phi) \cdot \bar{K}(\rho, \phi) - \sqrt{g}(\rho, \phi) \frac{\partial \bar{K}}{\partial \rho}(\rho, \phi).$$

We note that the term $\partial \bar{K} / \partial \rho(\rho, \phi)$ makes sense even for $\rho = 0$; it is just a directional derivative of \bar{K} , which is a C^∞ function on M_p . Consequently, $\partial \bar{K} / \partial \rho(\rho, \phi)$ approaches a limit as $\rho \rightarrow 0$. From (6) we thus obtain

$$\begin{aligned}
 (7) \quad \frac{\partial^3 \sqrt{g}}{\partial \rho^3}(0, \phi) &= \lim_{\rho \rightarrow 0} \frac{\partial^3 \sqrt{g}}{\partial \rho^3}(\rho, \phi) \\
 &= -K(p), \quad \text{using (2) and (3), and noting that} \\
 &\quad \bar{K}(0, \phi) = K(p).
 \end{aligned}$$

Using (2), (3), (5), (7), we now have the Taylor polynomial expansion

$$\boxed{\sqrt{g}(\rho, \phi) = \rho - \frac{K(p)\rho^3}{6} + o(\rho^3).}$$

[There is still a technical detail which must be taken care of. We actually want to know that the remainder

$$R(\rho) = \sqrt{g}(\rho, \phi) - \rho + \frac{K(p)\rho^3}{6}$$

satisfies

$$\lim_{\rho \rightarrow 0} \frac{R(\rho)}{\rho^3} \rightarrow 0 \quad \text{uniformly in } \phi.$$

This can be seen from the proof that the Taylor polynomial approximates the function; the proof involves L'Hôpital's Theorem, which in turn depends on the Cauchy mean value theorem (*Calculus*, pg. 201), whose role will be made explicit. We have

$$\begin{aligned}
 \lim_{\rho \rightarrow 0} \frac{\sqrt{g}(\rho, \phi) - \rho + K(p)\rho^3/6}{\rho^3} &= \lim_{\rho \rightarrow 0} \frac{\frac{\partial \sqrt{g}}{\partial \rho}(\bar{\rho}, \phi) - 1 + K(p)\bar{\rho}^2/2}{3\bar{\rho}^2} \\
 &\quad 0 < \bar{\rho} < \rho, \text{ by the} \\
 &\quad \text{Cauchy mean value theorem} \\
 &= \lim_{\rho \rightarrow 0} \frac{\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\bar{\bar{\rho}}, \phi) + K(p)\bar{\bar{\rho}}}{6\bar{\bar{\rho}}} \quad 0 < \bar{\bar{\rho}} < \bar{\rho} \\
 &= \lim_{\rho \rightarrow 0} \frac{\frac{\partial^3 \sqrt{g}}{\partial \rho^3}(\bar{\bar{\bar{\rho}}}, \phi) + K(p)}{6} \quad 0 < \bar{\bar{\bar{\rho}}} < \bar{\bar{\rho}} \\
 &= 0,
 \end{aligned}$$

with the final equality coming from (7).]

10. PROPOSITION (BERTRAND AND PUISEUX; 1848). Let $C(\rho)$ be the circumference of the “geodesic circle” of radius ρ around $p \in M$, consisting of the endpoints of geodesic segments of length ρ which start at p . Then

$$K(p) = \lim_{\rho \rightarrow 0} 3 \cdot \frac{2\pi\rho - C(\rho)}{\pi\rho^3}.$$

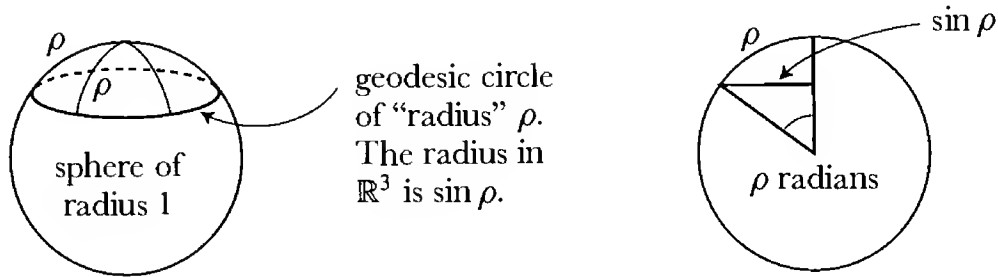
PROOF. Clearly

$$\begin{aligned} C(\rho) &= \int_0^{2\pi} \left\| \frac{\partial}{\partial \phi} \Big|_{\exp(\rho, \phi)} \right\| d\phi = \int_0^{2\pi} \sqrt{g}(\rho, \phi) d\phi \\ &= \int_0^{2\pi} \left(\rho - \frac{K(p)\rho^3}{6} \right) d\phi + \int_0^{2\pi} o(\rho^3) d\phi \\ &= 2\pi \left(\rho - \frac{K(p)\rho^3}{6} \right) + o(\rho^3). \end{aligned}$$

So

$$K(p) = 3 \cdot \frac{2\pi\rho - C(\rho)}{\pi\rho^3} + \frac{o(\rho^3)}{\rho^3}. \quad \blacklozenge$$

According to Proposition 10, on a surface of positive curvature, like a sphere, geodesic circles are always “too small”, while on surfaces of negative curvature,



they are always “too large”. Notice that Proposition 10 gives another interpretation of curvature totally in terms of the Riemannian metric on M . There is yet another formula of the same type.

11. PROPOSITION (DIQUET; 1848). Let $A(\rho)$ be the area enclosed by the geodesic circle of radius ρ around $p \in M$. Then

$$K(p) = \lim_{\rho \rightarrow 0} 12 \cdot \frac{\pi\rho^2 - A(\rho)}{\pi\rho^4}.$$

PROOF. Clearly

$$\begin{aligned}
 A(\rho) &= \int_{\substack{\text{inclosed} \\ \text{region}}} dV = \int_0^{2\pi} \int_0^\rho \sqrt{g}(\rho, \phi) d\rho d\phi \\
 &= \int_0^{2\pi} \int_0^\rho \left(\rho - \frac{K(p)\rho^3}{6} \right) d\rho d\phi + \int_0^{2\pi} \int_0^\rho o(\rho^3) d\rho d\phi \\
 &= 2\pi \left(\frac{\rho^2}{2} - \frac{K(p)\rho^4}{24} \right) + o(\rho^4),
 \end{aligned}$$

which easily yields the desired result. ♦

CHAPTER 4

THE CURVATURE OF HIGHER DIMENSIONAL MANIFOLDS

A. AN INAUGURAL LECTURE

On June 10, 1854 the faculty of Göttingen University heard a lecture entitled *Über die Hypothesen, welche der Geometrie zu Grunde liegen* (On the Hypotheses which lie at the Foundations of Geometry). This lecture was delivered by Georg Friedrich Bernhard Riemann, who had been born just a year before Gauss' paper of 1827. Although the lecture was not published until 1866, the ideas contained within it proved to be the most influential in the entire history of differential geometry. To be sure, mathematicians had not neglected the study of surfaces in the meantime; in fact, Gauss' work had inspired a tremendous amount of work along these lines. But the results obtained in those years can all be proved with much greater ease after we have followed the long series of developments initiated by the turning point in differential geometry which Riemann's lecture provided.

A short account of the life and character of Riemann can be found in the biography by Dedekind* which is included in Riemann's collected works (published by Dover). His interest in many fields of mathematical physics, together with a demand for perfection in all he did, delayed until 1851 the submission of his doctoral dissertation *Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse* (Foundations for a general theory of functions of a complex variable). Gauss' official report to the Philosophical Faculty of the University of Göttingen stated "The dissertation submitted by Herr Riemann offers convincing evidence of the author's thorough and penetrating investigations in those parts of the subject treated in the dissertation, of a creative, active truly mathematical mind, and of a gloriously fertile originality."

Riemann was now qualified to seek the position of Privatdocent (a lecturer who received no salary, but was merely forwarded fees paid by those students

* Even for those who can only plod through German, this is preferable to the account in E. T. Bell's *Men of Mathematics*, which is hardly more than a translation of Dedekind, written in a racy style and interlarded with supercilious remarks of questionable taste.

who elected to attend his lectures). To attain this position he first had to submit an “inaugural paper” (Habilitationsschrift). Again there were delays, and it was not until the end of 1853 that Riemann submitted the Habilitationsschrift, *Über die Darstellbarkeit einer Function durch eine trigonometrische Reihe* (On the representability of a function by a trigonometric series). Now Riemann still had to give a probationary inaugural lecture on a topic chosen by the faculty, from a list of three proposed by the candidate. The first two topics which Riemann submitted were ones on which he had already worked, and he had every reason to expect that one of these two would be picked; for the third topic he chose the foundations of geometry. Contrary to all traditions, Gauss passed over the first two, and picked instead the third, in which he had been interested for years. At this time Riemann was also investigating the connection between electricity, magnetism, light, and gravitation, in addition to acting as an assistant in a seminar on mathematical physics. The strain of carrying out another major investigation, aggravated perhaps by the hardships of poverty, brought on a temporary breakdown. However, Riemann soon recovered, disposed of some other work which had to be completed, and then finished his inaugural lecture in about seven more weeks.

Riemann hoped to make his lecture intelligible even to those members of the faculty who knew little mathematics. Consequently, hardly any formulas appear and the analytic investigations are completely suppressed. Although Dedekind describes the lecture as a masterpiece of exposition, it is questionable how many of the faculty comprehended it. In making the following translation,* I was aided by the fact that I already had some idea what the mathematical results were supposed to be. The uninitiated reader will probably experience a great deal of difficulty merely understanding what Riemann is trying to say (the proofs of Riemann’s assertions are spread out over the next several chapters). We can be sure, however, that one member of the faculty appreciated Riemann’s work. Dedekind tells us that Gauss sat at the lecture “which surpassed all his expectations, in the greatest astonishment, and on the way back from the faculty meeting he spoke to Wilhelm Weber, with the greatest appreciation, and with an excitement rare for him, about the depth of the ideas presented by Riemann”.

*The original is contained, of course, in Riemann’s collected works. Two English translations are readily available, one in Volume 2 of Smith’s *Source Book in Mathematics* (Dover), and one in Clifford’s *Mathematical Papers* (Chelsea).

On the Hypotheses which lie at The Foundations of Geometry

Plan of the Investigation.

As is well known, geometry presupposes the concept of space, as well as assuming the basic principles for constructions in space. It gives only nominal definitions of these things, while their essential specifications appear in the form of axioms. The relationship between these presuppositions is left in the dark; we do not see whether, or to what extent, any connection between them is necessary, or *a priori* whether any connection between them is even possible.

From Euclid to Legendre, the most famous of the modern reformers of geometry, this darkness has been dispelled neither by the mathematicians nor by the philosophers who have concerned themselves with it. This is undoubtedly because the general concept of multiply extended quantities, which includes spatial quantities, remains completely unexplored. I have therefore first set myself the task of constructing the concept of a multiply extended quantity from general notions of quantity. It will be shown that a multiply extended quantity is susceptible of various metric relations, so that Space constitutes only a special case of a triply extended quantity. From this however it is a necessary consequence that the theorems of geometry cannot be deduced from general notions of quantity, but that those properties which distinguish Space from other conceivable triply extended quantities can only be deduced from experience. Thus arises the problem of seeking out the simplest data from which the metric relations of Space can be determined, a problem which by its very nature is not completely determined, for there may be several systems of simple data which suffice to determine the metric relations of Space; for the present purposes, the most important system is that laid down as a foundation of geometry by Euclid. These data are — like all data — not logically necessary, but only of empirical certainty, they are hypotheses; one can therefore investigate their likelihood, which is certainly very great within the bounds of observation, and afterwards decide upon the legitimacy of extending them beyond the bounds of observation, both in the direction of the immeasurably large, and in the direction of the immeasurably small.

I. Concept of an n fold extended quantity.

In proceeding to attempt the solution of the first of these problems, the development of the concept of multiply extended quantity, I feel particularly entitled to request an indulgent hearing, as I am little practiced in these tasks

of a philosophical nature where the difficulties lie more in the concepts than in the construction, and because I could not make use of any previous studies, except for some very brief hints on the subject which Privy Councilor Gauss has given in his second memoir on Biquadratic Residues, in the *Gottingen Gelehrte Anzeige* and in the *Gottingen Jubilee-book*, and some philosophical researches of Herbart.

1.

Notions of quantity are possible only when there already exists a general concept which admits particular instances. These instances form either a continuous or a discrete manifold, depending on whether or not a continuous transition of instances can be found between any two of them; individual instances are called points in the first case and elements of the manifold in the second. Concepts whose particular instances form a discrete manifold are so numerous that some concept can always be found, at least in the more highly developed languages, under which any given collection of things can be comprehended (and consequently, in the study of discrete quantities, mathematicians could unhesitatingly proceed from the principle that given objects are to be regarded as all of one kind). On the other hand, opportunities for creating concepts whose instances form a continuous manifold occur so seldom in everyday life that color and the position of sensible objects are perhaps the only simple concepts whose instances form a multiply extended manifold. More frequent opportunities for creating and developing these concepts first occur in higher mathematics.

Particular portions of a manifold, distinguished by a mark or by a boundary, are called quanta. Their quantitative comparison is effected in the case of discrete quantities by counting, in the case of continuous quantities by measurement. Measuring involves the superposition of the quantities to be compared; it therefore requires a means of transporting one quantity to be used as a standard for the others. Otherwise, one can compare two quantities only when one is a part of the other, and then only as to "more" or "less", not as to "how much". The investigations which can be carried out in this case form a general division of the science of quantity, independent of measurement, where quantities are regarded, not as existing independent of position and not as expressible in terms of a unit, but as regions in a manifold. Such investigations have become a necessity for several parts of mathematics, e.g., for the treatment of many-valued analytic functions, and the dearth of such studies is one of the principal reasons why the celebrated theorem of Abel and the contributions of Lagrange, Pfaff and Jacobi to the general theory of differential equations have remained unfruitful for so long. From this portion of the science of extended quantity,

a portion which proceeds without any further assumptions, it suffices for the present purposes to emphasize two points, which will make clear the essential characteristic of an n fold extension. The first of these concerns the generation of the concept of a multiply extended manifold, the second involves reducing position fixing in a given manifold to numerical determinations.

2.

In a concept whose instances form a continuous manifold, if one passes from one instance to another in a well-determined way, the instances through which one has passed form a simply extended manifold, whose essential characteristic is, that from any point in it a continuous movement is possible in only two directions, forwards and backwards. If one now imagines that this manifold passes to another, completely different one, and once again in a well-determined way, that is, so that every point passes to a well-determined point of the other, then the instances form, similarly, a doubly extended manifold. In a similar way, one obtains a triply extended manifold when one imagines that a doubly extended one passes in a well-determined way to a completely different one, and it is easy to see how one can continue this construction. If one considers the process as one in which the objects vary, instead of regarding the concept as fixed, then this construction can be characterized as a synthesis of a variability of $n + 1$ dimensions from a variability of n dimensions and a variability of one dimension.

3.

I will now show, conversely, how one can break up a variability, whose boundary is given, into a variability of one dimension and a variability of lower dimension. One considers a piece of a manifold of one dimension — with a fixed origin, so that points of it may be compared with one another — varying so that for every point of the given manifold it has a definite value, continuously changing with this point. In other words, we take within the given manifold a continuous function of position, which, moreover, is not constant on any part of the manifold. Every system of points where the function has a constant value then forms a continuous manifold of fewer dimensions than the given one. These manifolds pass continuously from one to another as the function changes; one can therefore assume that they all emanate from one of them, and generally speaking this will occur in such a way that every point of the first passes to a definite point of any other; the exceptional cases, whose investigation is important, need not be considered here. In this way, the determination of position in the given manifold is reduced to a numerical determination and to the determination of

position in a manifold of fewer dimensions. It is now easy to show that this manifold has $n - 1$ dimensions, if the given manifold is an n fold extension. By an n time repetition of this process, the determination of position in an n fold extended manifold is reduced to n numerical determinations, and therefore the determination of position in a given manifold is reduced, whenever this is possible, to a finite number of numerical determinations. There are, however, also manifolds in which the fixing of position requires not a finite number, but either an infinite sequence or a continuous manifold of numerical measurements. Such manifolds form, e.g., the possibilities for a function in a given region, the possible shapes of a solid figure, etc.

II. Metric relations of which a manifold of n dimensions is susceptible, on the assumption that lines have a length independent of their configuration, so that every line can be measured by every other.

Now that the concept of an n fold extended manifold has been constructed, and its essential characteristic has been found in the fact that position fixing in the manifold can be reduced to n numerical determinations, there follows, as the second of the problems proposed above, an investigation of the metric relations of which such a manifold is susceptible, and of the conditions which suffice to determine them. These metric relations can be investigated only in abstract terms, and their interdependence exhibited only through formulas. Under certain assumptions, however, one can resolve them into relations which are individually capable of geometric representation, and in this way it becomes possible to express the results of calculation geometrically. Thus, although an abstract investigation with formulas certainly cannot be avoided, the results can be presented in geometric garb. The foundations of both parts of the question are contained in the celebrated treatise of Privy Councilor Gauss on curved surfaces.

1.

Measurement requires an independence of quantity from position, which can occur in more than one way. The hypothesis which first presents itself, and which I shall develop here, is just this, that the length of lines is independent of their configuration, so that every line can be measured by every other. If position-fixing is reduced to numerical determinations, so that the position of a point in the given n fold extended manifold is expressed by n varying quantities x_1, x_2, x_3 , and so forth up to x_n , then specifying a line amounts to giving the quantities x as functions of one variable. The problem then is, to set up a mathematical

expression for the length of a line, for which purpose the quantities x must be thought of as expressible in units. I will treat this problem only under certain restrictions, and I first limit myself to lines in which the ratios of the quantities dx — the increments in the quantities x — vary continuously; one can then regard the lines as broken up into elements within which the ratios of the quantities dx may be considered to be constant, and the problem then reduces to setting up a general expression for the line element ds at every point, an expression which will involve the quantities x and the quantities dx . I assume, secondly, that the length of the line element remains unchanged, up to first order, when all the points of this line element suffer the same infinitesimal displacement, whereby I simply mean that if all the quantities dx increase in the same ratio, the line element changes by the same ratio. Under these assumptions, the line element can be an arbitrary homogeneous function of the first degree in the quantities dx which remains the same when all the quantities dx change sign, and in which the arbitrary constants are functions of the quantities x . To find the simplest cases, I first seek an expression for the $(n - 1)$ fold extended manifolds which are everywhere equidistant from the origin of the line element, i.e., I seek a continuous function of position which distinguishes them from one another. This must either decrease or increase in all directions from the origin; I will assume that it increases in all directions and therefore has a minimum at the origin. Then if its first and second differential quotients are finite, the first order differential must vanish and the second order differential cannot be negative; I assume that it is always positive. This differential expression of the second order remains constant if ds remains constant and increases quadratically when the quantities dx , and thus also ds , all increase in the same ratio; it is therefore $= \text{constant} \cdot ds^2$ and consequently $ds =$ the square root of an everywhere positive homogeneous function of the second degree in the quantities dx , in which the coefficients are continuous functions of the quantities x . In Space, if one expresses the location of a point by rectilinear coordinates, then $ds = \sqrt{\Sigma(dx)^2}$; Space is therefore included in this simplest case. The next simplest case would perhaps include the manifolds in which the line element can be expressed as the fourth root of a differential expression of the fourth degree. Investigation of this more general class would actually require no essentially different principles, but it would be rather time consuming and throw proportionally little new light on the study of Space, especially since the results cannot be expressed geometrically; I consequently restrict myself to those manifolds where the line element can be expressed by the square root of a differential expression of the second degree. One can transform such an expression into another similar one by substituting for the n independent variables, functions of n new independent variables.

However, one cannot transform any expression into any other in this way; for the expression contains $n\frac{n+1}{2}$ coefficients which are arbitrary functions of the independent variables; by the introduction of new variables one can satisfy only n conditions, and can therefore make only n of the coefficients equal to given quantities. There remain $n\frac{n-1}{2}$ others, already completely determined by the nature of the manifold to be represented, and consequently $n\frac{n-1}{2}$ functions of position are required to determine its metric relations. Manifolds, like the Plane and Space, in which the line element can be brought into the form $\sqrt{\Sigma dx^2}$ thus constitute only a special case of the manifolds to be investigated here; they clearly deserve a special name, and consequently, these manifolds, in which the square of the lines element can be expressed as the sum of the squares of complete differentials, I propose to call flat. In order to survey the essential differences of the manifolds representable in the assumed form, it is necessary to eliminate the features depending on the mode of presentation, which is accomplished by choosing the variable quantities according to a definite principle.

2.

For this purpose, one constructs the system of shortest lines emanating from a given point; the position of an arbitrary point can then be determined by the initial direction of the shortest line in which it lies, and its distance, in this line, from the initial point. It can therefore be expressed by the ratios of the quantities dx^0 , i.e., the quantities dx at the origin of this shortest line, and by the length s of this line. In place of the dx^0 one now introduces linear expressions $d\alpha$ formed from them in such a way that the initial value of the square of the line element will be equal to the sum of the squares of these expressions, so that the independent variables are: the quantity s and the ratio of the quantities $d\alpha$. Finally, in place of the $d\alpha$ choose quantities x_1, x_2, \dots, x_n proportional to them, but such that the sum of their squares equals s^2 . If one introduces these quantities, then for infinitely small values of x the square of the line element = Σdx^2 , but the next order term in its expansion equals a homogeneous expression of the second degree in the $n\frac{n-1}{2}$ quantities $(x_1 dx_2 - x_2 dx_1), (x_1 dx_3 - x_3 dx_1), \dots$, and is consequently an infinitely small quantity of the fourth order, so that one obtains a finite quantity if one divides it by the square of the infinitely small triangle at whose vertices the variables have the values $(0, 0, 0, \dots), (x_1, x_2, x_3, \dots), (dx_1, dx_2, dx_3, \dots)$. This quantity remains the same as long as the quantities x and dx are contained in the same binary linear forms, or as long as the two shortest lines from the initial point to x and from the initial point to dx remain in the same surface element, and therefore depends only on the position and direction

of that element. It obviously = zero if the manifold in question is flat, i.e., if the square of the line element is reducible to Σdx^2 , and can therefore be regarded as the measure of deviation from flatness in this surface direction at this point. When multiplied by $-\frac{3}{4}$ it becomes equal to the quantity which Privy Councilor Gauss has called the curvature of a surface. Previously, $n^{\frac{n-1}{2}}$ functions of position were found necessary in order to determine the metric relations of an n fold extended manifold representable in the assumed form; hence if the curvature is given in $n^{\frac{n-1}{2}}$ surface directions at every point, then the metric relations of the manifold may be determined, provided only that no identical relations can be found between these values, and indeed in general this does not occur. The metric relations of these manifolds, in which the line element can be represented as the square root of a differential expression of the second degree, can thus be expressed in a way completely independent of the choice of the varying quantities. A similar path to the same goal could also be taken in those manifolds in which the line element is expressed in a less simple way, e.g., by the fourth root of a differential expression of the fourth degree. The line element in this more general case would not be reducible to the square root of a quadratic sum of differential expressions, and therefore in the expression for the square of the line element the deviation from flatness would be an infinitely small quantity of the second dimension, whereas for the other manifolds it was an infinitely small quantity of the fourth dimension. This peculiarity of the latter manifolds therefore might well be called plainness in the smallest parts. For present purposes, however, the most important peculiarity of these manifolds, on whose account alone they have been examined here, is this, that the metric relations of the doubly extended ones can be represented geometrically by surfaces and those of the multiply extended ones can be reduced to those of the surfaces contained within them, which still requires a brief discussion.

3.

In the conception of surfaces, the inner metric relations, which involve only the lengths of paths within them, are always bound up with the way the surfaces are situated with respect to points outside them. We may, however, abstract from external relations by considering deformations which leave the lengths of lines within the surfaces unaltered, i.e., by considering arbitrary bendings — without stretching — of such surfaces, and by regarding all surfaces obtained from one another in this way as equivalent. Thus, for example, arbitrary cylindrical or conical surfaces count as equivalent to a plane, since they can be formed from a plane by mere bending, under which the inner metric relations remain the same; and all theorems about the plane — hence all of planimetry

— retain their validity. On the other hand, they count as essentially different from the sphere, which cannot be transformed into the plane without stretching. According to the previous investigations, the inner metric relations at every point of a doubly extended quantity, if its line element can be expressed as the square root of a differential expression of the second degree, which is the case with surfaces, is characterized by the curvature. For surfaces, this quantity can be given a visual interpretation as the product of the two curvatures of the surface at this point, or by the fact that its product with an infinitely small triangle formed from shortest lines is, in proportion to the radius, half the excess of the sum of its angles over two right angles. The first definition would presuppose the theorem that the product of the two radii of curvatures is unaltered by mere bendings of a surface, the second, that at each point the excess over two right angles of the sum of the angles of any infinitely small triangle is proportional to its area. To give a tangible meaning to the curvature of an n fold extended manifold at a given point, and in a given surface direction through it, we first mention that a shortest line emanating from a point is completely determined if its initial direction is given. Consequently we obtain a certain surface if we prolong all the initial directions from the given point which lie in the given surface element, into shortest lines; and this surface has a definite curvature at the given point, which is equal to the curvature of the n fold extended manifold at the given point, in the given surface direction.

4.

Before applying these results to Space, it is still necessary to make some general considerations about flat manifolds, i.e., about manifolds in which the square of the line element can be represented as the sum of squares of complete differentials.

In a flat n fold extended manifold the curvature in every direction, at every point, is zero; but according to the preceding investigation, in order to determine the metric relations it suffices to know that at each point the curvature is zero in $n \frac{n-1}{2}$ independent surface-directions. The manifolds whose curvature is everywhere $= 0$ can be considered as a special case of those manifolds whose curvature is everywhere constant. The common character of those manifolds whose curvature is constant may be expressed as follows: figures can be moved in them without stretching. For obviously figures could not be freely shifted and rotated in them if the curvature were not the same in all directions, at all points. On the other hand, the metric properties of the manifold are completely determined by the curvature; they are therefore exactly the same in all the directions around any one point as in the directions around any other, and thus

the same constructions can be effected starting from either; consequently, in the manifolds with constant curvature figures may be given any arbitrary position. The metric relations of these manifolds depend only on the value of the curvature, and it may be mentioned, as regards the analytic presentation, that if one denotes this value by α , then the expression for the line element can be put in the form

$$\frac{1}{1 + \frac{\alpha}{4} \Sigma x^2} \sqrt{\Sigma dx^2}$$

5.

The consideration of *surfaces* with constant curvature may serve for a geometric illustration. It is easy to see that the surfaces whose curvature is positive can always be rolled onto a sphere whose radius is the reciprocal of the curvature; but in order to survey the multiplicity of these surfaces, let one of them be given the shape of a sphere, and the others the shape of surfaces of rotation which touch it along the equator. The surfaces with greater curvature than the sphere will then touch the sphere from inside and take a form like the portion of the surface of a ring, which is situated away from the axis; they could be rolled upon zones of spheres with smaller radii, but would go round more than once. Surfaces with smaller positive curvature are obtained from spheres of larger radii by cutting out a portion bounded by two great semi-circles, and bringing together the cut-lines. The surface of curvature zero will be a cylinder standing on the equator; the surfaces with negative curvature will touch this cylinder from outside and be formed like the part of the surface of a ring which is situated near the axis. If one regards these surfaces as possible positions for pieces of surface moving in them, as Space is for bodies, then pieces of surface can be moved in all these surfaces without stretching. The surfaces with positive curvature can always be so formed that pieces of surface can even be moved arbitrarily without bending, namely as spherical surfaces, but those with negative curvature cannot. Aside from this independence of position for surface pieces, in surfaces with zero curvature there is also an independence of position for directions, which does not hold in the other surfaces.

III. Applications to Space.

1.

Following these investigations into the determination of the metric relations of an n fold extended quantity, the conditions may be given which are sufficient and necessary for determining the metric relations of Space, if we assume

beforehand the independence of lines from configuration and the possibility of expressing the line element as the square root of a second order differential expression, and thus flatness in the smallest parts.

First, these conditions may be expressed by saying that the curvature at every point equals zero in three surface directions, and thus the metric relations of Space are implied if the sum of the angles of a triangle always equals two right angles.

But secondly, if one assumes with Euclid not only the existence of lines independently of configuration, but also of bodies, then it follows that the curvature is everywhere constant, and the angle sum in all triangles is determined if it is known in one.

In the third place, finally, instead of assuming the length of lines to be independent of place and direction, one might assume that their length and direction is independent of place. According to this conception, changes or differences in position are complex quantities expressible in three independent units.

2.

In the course of the previous considerations, the relations of extension or regionality were first distinguished from the metric relations, and it was found that different metric relations were conceivable along with the same relations of extension; then systems of simple metric specifications were sought by means of which the metric relations of Space are completely determined, and from which all theorems about it are a necessary consequence. It remains now to discuss the question how, to what degree, and to what extent these assumptions are borne out by experience. In this connection there is an essential difference between mere relations of extension and metric relations, in that among the first, where the possible cases form a discrete manifold, the declarations of experience are to be sure never completely certain, but they are not inexact, while for the second, where the possible cases form a continuous manifold, every determination from experience always remains inexact — be the probability ever so great that it is nearly exact. This circumstance becomes important when these empirical determinations are extended beyond the limits of observation into the immeasurably large and the immeasurably small; for the latter may obviously become ever more inexact beyond the boundary of observation, but not so the former.

When constructions in Space are extended into the immeasurably large, unboundedness is to be distinguished from infinitude; one belongs to relations of extension, the other to metric relations. That Space is an unbounded triply

extended manifold is an assumption which is employed for every apprehension of the external world, by which at every moment the domain of actual perception is supplemented, and by which the possible locations of a sought for object are constructed; and in these applications it is continually confirmed. The unboundedness of space consequently has a greater empirical certainty than any experience of the external world. But its infinitude does not in any way follow from this; quite to the contrary, Space would necessarily be finite if one assumed independence of bodies from position, and thus ascribed to it a constant curvature, as long as this curvature had ever so small a positive value. If one prolonged the initial directions lying in a surface direction into shortest lines, one would obtain an unbounded surface with constant positive curvature, and thus a surface which in a flat triply extended manifold would take the form of a sphere, and consequently be finite.

3.

Questions about the immeasurably large are idle questions for the explanation of Nature. But the situation is quite different with questions about the immeasurably small. Upon the exactness with which we pursue phenomena into the infinitely small, does our knowledge of their causal connections essentially depend. The progress of recent centuries in understanding the mechanisms of Nature depends almost entirely on the exactness of construction which has become possible through the invention of the analysis of the infinite and through the simple principles discovered by Archimedes, Galileo, and Newton, which modern physics makes use of. By contrast, in the natural sciences where the simple principles for such constructions are still lacking, to discover causal connections one pursues phenomenon into the spatially small, just so far as the microscope permits. Questions about the metric relations of Space in the immeasurably small are thus not idle ones.

If one assumes that bodies exist independently of position, then the curvature is everywhere constant, and it then follows from astronomical measurements that it cannot be different from zero; or at any rate its reciprocal must be an area in comparison with which the range of our telescopes can be neglected. But if such an independence of bodies from position does not exist, then one cannot draw conclusions about metric relations in the infinitely small from those in the large; at every point the curvature can have arbitrary values in three directions, provided only that the total curvature of every measurable portion of Space is not perceptibly different from zero. Still more complicated relations can occur if the line element cannot be represented, as was presupposed, by the square root of a differential expression of the second degree. Now it seems that

the empirical notions on which the metric determinations of Space are based, the concept of a solid body and that of a light ray, lose their validity in the infinitely small; it is therefore quite definitely conceivable that the metric relations of Space in the infinitely small do not conform to the hypotheses of geometry; and in fact one ought to assume this as soon as it permits a simpler way of explaining phenomena.

The question of the validity of the hypotheses of geometry in the infinitely small is connected with the question of the basis for the metric relations of Space. In connection with this question, which may indeed still be ranked as part of the study of Space, the above remark is applicable, that in a discrete manifold the principle of metric relations is already contained in the concept of the manifold, but in a continuous one it must come from something else. Therefore, either the reality underlying Space must form a discrete manifold, or the basis for the metric relations must be sought outside it, in binding forces acting upon it.

An answer to these questions can be found only by starting from that conception of phenomena which has hitherto been approved by experience, for which Newton laid the foundation, and gradually modifying it under the compulsion of facts which cannot be explained by it. Investigations like the one just made, which begin from general concepts, can serve only to insure that this work is not hindered by unduly restricted concepts, and that progress in comprehending the connection of things is not obstructed by traditional prejudices.

This leads us away into the domain of another science, the realm of physics, into which the nature of the present occasion does not allow us to enter.

B. WHAT DID RIEMANN SAY?

Upon a first reading, Riemann's lecture may appear to have almost no mathematical content. But this is only because the analytic investigations, which occur in Part II, have been drastically condensed, while Part I explains, in general philosophical terms, important mathematical concepts which succeeding generations of investigators were eventually able to express with mathematical precision; finally, Part III of the lecture deals with applications of the mathematical discoveries to questions in physics, a process which is perhaps not yet complete.

In this commentary on Riemann's lecture, we will follow closely the order of Riemann's exposition, referring often to the various sections (1, 2, etc.) within each part (I, II, III). It should not be expected that all details will be cleared up, even in the remaining portions of this chapter, for the complete consideration of Riemann's ideas will occupy several of the succeeding chapters. Consequently, the remaining parts of Chapter 4 may be the hardest reading encountered in either of the two volumes of these notes. Nevertheless, we hope that in the end a clear view of all these ideas will be obtained.

In the "Plan of the Investigation", Riemann begins by accounting for the confusion over the status of non-Euclidean geometry, which at this time was still not completely accepted. In 1829, Lobachevsky and Bolyai had independently constructed a system of geometry which began by assuming that through a point not on a line there was more than one line parallel to it (as opposed to the assumption that there is only one parallel line, which is equivalent to Euclid's Fifth Postulate); but it was still supposed by some that contradictions in this system would eventually be found.

Riemann attributes the difficulties encountered in the study of non-Euclidean geometry to the fact that geometers had never separated what we would call the topological properties of space from its metric properties; in the axiomatic development of geometry, even the notion of space itself is undefined, and its properties are developed through the axioms.

Riemann proposes to distinguish the metric properties from the topological properties, and promises that we will discover how different metric structures can be put on the triply extended quantity which constitutes Space, so that one cannot possibly expect to deduce the parallel postulate of Euclid from topological considerations alone. This implies that experimental data must be used to determine what metric properties Space actually has, and raises the question which data we should seek, and what we can expect to say about the regions of Space too distant, or too small, to be investigated experimentally.

In Part I, “Concept of an n fold extended quantity”, Riemann is clearly trying to define a manifold.

It is impossible to tell from this lecture, intended for non-mathematicians, how far Riemann had advanced toward the precise solution of this problem, and whether he had any way of expressing concretely the notion of a metric or topological space, which is essentially prerequisite to the definition of a manifold. However, it is quite obvious that the notion was thoroughly clear in his own mind and that he recognized that manifolds were characterized by the fact that they are locally like n -dimensional Euclidean space. It is also clear that he understood the importance of infinite dimensional spaces, such as the set of all real-valued functions on a space (it is interesting that quite recently some of these infinite dimensional spaces have been given the structure of “infinite dimensional manifolds”, and differential geometric methods have been applied to them with great success).

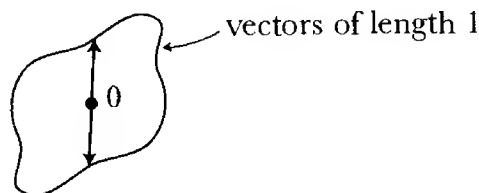
Part II contains nearly all the mathematical results, and the discussion of this Part will take up most of the present chapter.

The difficulties in Part II begin right with the title, “Metric relations of which a manifold of n dimensions is susceptible, on the assumption that lines have a length independent of their configuration, so that every line can be measured by every other”. To understand what Riemann means, it is necessary to recall the process by which lengths are assigned to curves in the plane or 3-dimensional space of analytic geometry. In this case, we begin with the notion of distance between pairs of points, which amounts to saying that we first assign a length to *straight* lines; the length of other lines is then defined as the least upper bound of inscribed curves made up of straight lines, a process which can be reduced to integration. In this method of assigning lengths to curves, it may be said that all curves are measured by means of straight lines.

By contrast, Riemann proposes to consider a uniform method of assigning lengths to all curves in a manifold, a method which does not depend on first distinguishing a particular class of curves. This is to be done by measuring the lengths of tangent vectors, so that the lengths of curves can be defined by an integral (a restriction to C^1 curves is first indicated). Riemann assumes that this “length” function f is continuous on each tangent space and also positive homogeneous—the length $f(\lambda v)$ of λv is $|\lambda|$ times the length $f(v)$ of v .

Now there are many kinds of positive homogeneous functions on a finite dimensional vector space; any subset of the vector space which is symmetric with respect to the origin, and intersects each ray through 0 just once, can be used as the set of vectors of length 1. Riemann notes that the partial derivative of f^2 (with respect to some basis of the tangent space M_p) must vanish at

$0 \in M_p$, and that the matrix of second order partial derivatives is positive semi-



definite. He then assumes, as the simplest possibility, that it is actually positive definite. This means that f can be expressed as $\sqrt{\sum g_{ij}(p) dx^i \cdot dx^j}$ for certain numbers $g_{ij}(p)$. An assignment to each tangent space M_p of such a norm, or more precisely the inner product from which it comes, is, of course, what we now call a Riemannian metric on the manifold M .

Riemann points out that it is merely to save time, and to allow geometric descriptions of the results, that he restricts his attention to the special case. Certain more general cases, though not the most general of all, were investigated by Finsler in his thesis (1918), and are now known as Finsler metrics; it seems clear, however, that Riemann must have already known the basic facts about these more general metrics (some information on Finsler metrics is given in the Addendum).

Having restricted his attention to "Riemannian manifolds", Riemann now asks the crucial question: when does the introduction of a new coordinate system change the metric $\sum g_{ij} dy^i \otimes dy^j$ into some given metric $\sum a_{ij} dx^i \otimes dx^j$; in other words, when are two Riemannian manifolds locally isometric? Riemann here presents one of his famous "counting arguments", which enabled him to guess results that in some cases were not rigorously proved until a hundred years later. Riemann argues that the expression $\sum g_{ij} dx^i \otimes dx^j$ contains $n\frac{n+1}{2}$ functions (not n^2 , for $g_{ij} = g_{ji}$) while a new coordinate system involves only n functions, so that we can change only n of the g_{ij} , leaving $n\frac{n-1}{2}$ other functions which depend on the metric; consequently, Riemann argues, there should be some set of $n\frac{n-1}{2}$ functions which will determine the metric completely.

In section 2 of Part II, Riemann indicates how such functions are to be found. We are going to apply a standard technique for the study of differentiable functions—we examine the Taylor polynomials approximating the functions g_{ij} . If x is a coordinate system on M , with $x(p) = 0$, and the Riemannian metric is given by $\langle \ , \ \rangle = \sum g_{ij} dx^i \otimes dx^j$, then for the Taylor expansion of the

function $g_{ij} \circ x^{-1}: \mathbb{R}^n \rightarrow \mathbb{R}$ we have

$$\begin{aligned} g_{ij} \circ x^{-1}(t) &= (g_{ij} \circ x^{-1})(0) + \sum_{k=1}^n D_k(g_{ij} \circ x^{-1})(0)t^k \\ &\quad + \frac{1}{2} \sum_{k,l=1}^n D_{k,l}(g_{ij} \circ x^{-1})(0)t^k t^l + o(|t|^2). \end{aligned}$$

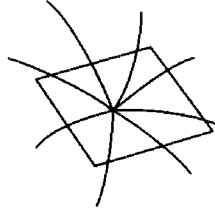
Hence on M we have

$$(*) \quad g_{ij} = g_{ij}(p) + \sum_{k=1}^n \frac{\partial g_{ij}}{\partial x^k}(p)x^k + \frac{1}{2} \sum_{k,l=1}^n \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}(p)x^k x^l + o(|x|^2),$$

where $o(|x|^2)$ denotes a function f on M such that

$$\lim_{q \rightarrow p} \frac{f(q)}{|x(q)|^2} = 0.$$

However, and this is the important device Riemann introduces, we will select a very special coordinate system around each point $p \in M$. We choose an orthonormal basis $X_1, \dots, X_n \in M_p$, and define a coordinate system $\chi: M_p \rightarrow \mathbb{R}^n$ on M_p by $\chi(\sum a^i X_i) = (a^1, \dots, a^n)$. Then we let x be the coordinate system $\chi \circ \exp^{-1}$. (This coordinate system is introduced at the very beginning



of section 2, but it takes a little work to decipher Riemann's description of it.)

The coordinate system x is not uniquely determined, for it depends on the choice of the orthonormal basis $X_1, \dots, X_n \in M_p$; but any two differ by an element of $O(n)$, so it will not be hard to take into account the way any of our results depend on this choice. These coordinate systems are called **Riemannian normal coordinates** at p . Notice that since $\exp_*: (M_p)_0 \rightarrow M_p$ is the identity (upon identifying $(M_p)_0$ with M_p), we have

$$\left. \frac{\partial}{\partial x^i} \right|_p = \exp_* X_i = X_i \in M_p.$$

We can quickly give some information about the first two terms in the expansion (*) of g_{ij} :

1. PROPOSITION. In a Riemannian normal coordinate system x at p we have

$$g_{ij}(p) = \delta_{ij}$$

$$\frac{\partial g_{ij}}{\partial x^k}(p) = 0.$$

PROOF. The first set of equations is clear, for

$$g_{ij}(p) = \left\langle \frac{\partial}{\partial x^i} \Big|_p, \frac{\partial}{\partial x^j} \Big|_p \right\rangle = \langle X_i, X_j \rangle = \delta_{ij}.$$

To prove the second set of equations, we recall the equations for a geodesic γ :

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0.$$

In Riemannian normal coordinates the geodesics through p are just $\exp \circ c$, where c is a straight line in M_p . This means that for all n -tuples (ξ^1, \dots, ξ^n) , the geodesics through p are the curves γ with $\gamma^k(t) = \xi^k t$. Hence

$$\sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \xi^i \xi^j = 0 \quad \text{for the geodesic } \gamma^k(t) = \xi^k t.$$

In particular, since $p = \gamma(0)$ is on all these geodesics, we have

$$\sum_{i,j=1}^n \Gamma_{ij}^k(p) \xi^i \xi^j = 0 \quad \text{for all } n\text{-tuples } (\xi^1, \dots, \xi^n).$$

This shows that all $\Gamma_{ij}^k(p)$ are 0: choosing all $\xi^\alpha = 0$ except $\xi^i = 1$ gives $\Gamma_{ii}^k = 0$; then choosing all $\xi^\alpha = 0$ except $\xi^i = \xi^j = 1$ gives

$$0 = \Gamma_{ij}^k(p) + \Gamma_{ji}^k(p) + \Gamma_{ii}^k(p) + \Gamma_{jj}^k(p) = 2\Gamma_{ij}^k(p).$$

It follows that

$$[ij, k] = \sum_{\alpha=1}^n g_{\alpha l} \Gamma_{ij}^l = 0 \quad \text{at } p.$$

Making use of equation (*) on pg. I.331, we have finally,

$$\frac{\partial g_{ij}}{\partial x^k} = [ik, j] + [jk, i] = 0 \quad \text{at } p. \quad \spadesuit$$

In view of Proposition 1, we can now use (*) to expand the squared norm $\| \|^2$ as

$$\begin{aligned} \| \|^2 &= \sum_{i,j=1}^n g_{ij} dx^i dx^j \\ &= \sum_{i=1}^n dx^i dx^i + \frac{1}{2} \sum_{i,j;k,l} \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}(p) x^k x^l dx^i dx^j + o(|x|^2). \end{aligned}$$

[This is an equation for tangent vectors near p , and $o(|x|^2)$ now denotes a function f on tangent vectors; in order to have

$$\lim_{q \rightarrow p} \frac{f(v_q)}{|x|^2} = 0,$$

we must restrict v_q to be of some bounded length.] Riemann's main assertion involves the term

$$\frac{1}{2} \sum_{i,j;k,l} \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}(p) x^k x^l dx^i dx^j = \sum_{i,j;k,l} c_{ij,kl} x^k x^l dx^i dx^j, \quad \text{say.}$$

Riemann asserts that there are numbers $C_{ij,kl}$ such that we can write

$$\sum_{i,j;k,l} c_{ij,kl} x^k x^l dx^i dx^j = \sum_{i,j;k,l} C_{ij,kl} (x^k dx^i - x^i dx^k) \cdot (x^l dx^j - x^j dx^l).$$

This assertion immediately suggests three questions—Why did Riemann suspect this was true? How did he prove it? What is its significance?

We will begin by giving a partial answer to the third of these questions. Notice that the equation in question is supposed to hold for all tangent vectors v at all points q in a neighborhood of p . Consequently, the numbers $dx^i(v)$ [and $x^i(q)$] can take on all [sufficiently small] values. The coordinate system x and the Riemannian metric $\langle \cdot, \cdot \rangle$ are used to obtain the n^4 numbers $c_{ij,kl} = \frac{1}{2} \partial^2 g_{ij} / \partial x^k \partial x^l(p)$; but beyond this, the above equation has nothing to do with the manifold at all. If we define a quadratic polynomial Q in $2n$ variables by

$$Q(X, Y) = Q(X_1, \dots, X_n, Y_1, \dots, Y_n) = \sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l,$$

then Riemann is asserting that this quadratic polynomial can be written as

$$Q(X, Y) = \sum_{i,j;k,l} C_{ij,kl} (X_i Y_k - X_k Y_i) \cdot (X_j Y_l - X_l Y_j).$$

To obtain the geometric consequences of this fact, we observe what it says when we select two vectors $v_p, w_p \in M_p$ and let $X_i = dx^i(v_p)$ and $Y_i = dx^i(w_p)$; denoting $Q(X, Y)$ by $Q(v_p, w_p)$, we have

$$\begin{aligned} Q(v_p, w_p) &= \sum_{i,j,k,l} c_{ij,kl} dx^i(v_p) dx^j(v_p) \cdot dx^k(w_p) dx^l(w_p) \\ &= \sum_{i,j,k,l} C_{ij,kl} [(dx^i \wedge dx^k)(v_p, w_p)] \cdot [(dx^j \wedge dx^l)(v_p, w_p)]. \end{aligned}$$

[We can also write

$$Q = \sum_{i,j,k,l} c_{ij,kl} dx^i dx^j \otimes dx^k dx^l = \sum_{i,j,k,l} C_{ij,kl} (dx^i \wedge dx^k) \cdot (dx^j \wedge dx^l),$$

a little more simply.] Now suppose $v'_p, w'_p \in M_p$ span the same subspace as v_p, w_p , so that we can write

$$\begin{aligned} v'_p &= a_{11}v_p + a_{21}w_p \\ w'_p &= a_{12}v_p + a_{22}w_p \end{aligned} \quad \det(a_{ij}) \neq 0.$$

The right side of the above equation for $Q(v_p, w_p)$ shows that

$$Q(v'_p, w'_p) = [\det(a_{ij})]^2 \cdot Q(v_p, w_p),$$

since each $dx^\alpha \wedge dx^\beta$ is multiplied by the factor $\det(a_{ij})$. If we use $\|v_p, w_p\|$ to denote the area of the parallelogram spanned by v_p and w_p , then we also have

$$\|v'_p, w'_p\|^2 = [\det(a_{ij})]^2 \cdot \|v_p, w_p\|^2.$$

Consequently,

$$\frac{Q(v'_p, w'_p)}{\|v'_p, w'_p\|^2} = \frac{Q(v_p, w_p)}{\|v_p, w_p\|^2}.$$

We therefore have a way of assigning a number to every 2-dimensional subspace of the tangent space at p . (Riemann sticks to the original quadratic function of the x^i and dx^i , which puts him in the position of having to divide by the squared area of a very strange triangle, with one vertex at x^i , and one at dx^i .)

It is easy to see that if we pick a different Riemannian normal coordinate system at p , then the resulting function on the 2-dimensional subspaces of M_p will be the same, for $Q(v_p, w_p) = Q(dx^1(v_p), \dots, dx^n(v_p), dx^1(w_p), \dots, dx^n(w_p))$ will change by $(\det B)^2$, where $B \in O(n)$, so that $\det B = \pm 1$. We will examine later the significance of this new function on 2-dimensional subspaces of the tangent space. For the present we take up the other questions—Why did Riemann think it was true, and how did he prove it?

Of course, an answer to the first question is not only doomed to be mere conjecture, but is always foolhardy to put forth, for there is no accounting for genius. The best suggestion I can offer is that the dependence of $Q(v_p, w_p)$ on the span of v_p and w_p alone is certainly an attractive one, and as we shall see later, in one special case which Riemann may have investigated first, the result appears in a rather natural way. It is also impossible to say for sure how Riemann proved the result, for his own investigations were never published. I have used the remarks by H. Weber in Riemann's collected works (pp. 405–409), as well as the commentary given by Herman Weyl in a special edition of Riemann's lecture. There are two parts to the proof, a purely algebraic one about quadratic functions, which determines what relations the numbers $c_{ij,kl}$ ought to satisfy, and an analytic one which establishes these relations.

For the algebraic part, we will be considering a quadratic function Q of $2n$ variables

$$Q(X, Y) = Q(X_1, \dots, X_n, Y_1, \dots, Y_n) = \sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l.$$

Note that for our Q we have

$$c_{ij,kl} = c_{ji,kl} = c_{ij,lk},$$

(using $g_{ij} = g_{ji}$, and $\partial^2/\partial x^k \partial x^l = \partial^2/\partial x^l \partial x^k$). If $A = (a_{ij})$ is a 2×2 matrix, we will use $A(X, Y)$ to denote the $2n$ -tuple

$$\begin{aligned} A(X, Y) &= (a_{11}X + a_{21}Y, a_{12}X + a_{22}Y) \\ &= (a_{11}X_1 + a_{21}Y_1, \dots, a_{11}X_n + a_{21}Y_n, a_{12}X_1 + a_{22}Y_1, \dots, a_{12}X_n + a_{22}Y_n). \end{aligned}$$

2. PROPOSITION. Let Q be a quadratic function of $2n$ variables,

$$Q(X, Y) = \sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l,$$

where

$$(1) \quad c_{ij,kl} = c_{ji,kl} = c_{ij,lk}.$$

Then

$$Q(A(X, Y)) = (\det A)^2 Q(X, Y)$$

for all 2×2 matrices A if and only if:

$$(2) \quad c_{ij,kl} = c_{kl,ij}$$

$$(3) \quad c_{li,jk} + c_{lj,ki} + c_{lk,ij} = 0.$$

PROOF. First of all, the equation $Q(A(X, Y)) = (\det A)^2 Q(X, Y)$ clearly holds for all 2×2 matrices A if and only if it holds for the non-singular ones, since both sides of the equation are continuous functions of A , and the non-singular matrices are dense.

Now it is well known that all non-singular 2×2 matrices can be written as a product of the matrices

$$\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

[This comes from the fact that any non-singular matrix can be obtained from the identity matrix by a sequence of elementary row operations, and every row operation may be accomplished by multiplying by one of the above matrices.] So our condition holds for all A if and only if it holds for the above matrices. We can disregard the last matrix, since

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For the matrix $A = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}$, the condition $Q(A(X, Y)) = (\det A)^2 Q(X, Y)$ becomes simply

$$Q(aX, Y) = a^2 Q(X, Y),$$

which is automatically true. The same result holds for the second matrix on our list, so all the conditions finally come down to

$$(a) \quad Q(Y, X) = Q(X, Y) \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$(b) \quad Q(X + Y, Y) = Q(X, Y) \quad A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Now equation (a) becomes

$$\sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l = \sum_{i,j;k,l} c_{ij,kl} Y_i Y_j X_k X_l.$$

Since this must be a polynomial identity, we obtain (2) immediately, by looking at the coefficient of $X_i X_j Y_k Y_l$ on both sides.

Equation (b) becomes

$$\sum_{i,j;k,l} c_{ij,kl} (X_i + Y_i)(X_j + Y_j) Y_k Y_l = \sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l,$$

or

$$\sum_{i,j;k,l} c_{ij,kl} [X_i Y_j Y_k Y_l + X_j Y_i Y_k Y_l + Y_i Y_j Y_k Y_l] = 0.$$

Letting $X = 0$, we obtain

$$(b1) \quad \sum_{i,j;k,l} c_{ij,kl} Y_i Y_j Y_k Y_l = 0$$

and then, in consequence,

$$(b2) \quad \sum_{i,j;k,l} c_{ij,kl} [X_i Y_j Y_k Y_l + X_j Y_i Y_k Y_l] = 0.$$

On the other hand, (b2) implies (b1), so (b) is equivalent to (b2) alone. Finally, since $c_{ij,kl} = c_{ji,kl}$, equation (b2) is equivalent to

$$(b3) \quad \sum_{i,j;k,l} c_{ij,kl} X_i Y_j Y_k Y_l = 0.$$

Looking at the coefficient of a particular $X_i Y_j Y_k Y_l$ we obtain

$$c_{ij,kl} + c_{ij,lk} + c_{ik,jl} + c_{ik,lj} + c_{il,jk} + c_{il,kj} = 0.$$

Using the symmetry with respect to the last two indices, this is equivalent to equation (3). ♦

3. PROPOSITION. A quadratic function

$$Q(X, Y) = \sum_{i,j;k,l} c_{ij,kl} X_i X_j Y_k Y_l$$

with

$$(1) \quad c_{ij,kl} = c_{ji,kl} = c_{ij,kl}$$

satisfies the two equivalent conditions of Proposition 2 if and only if it can be written as

$$Q(X, Y) = \sum_{i,j;k,l} C_{ij,kl} (X_i Y_k - X_k Y_i) \cdot (X_j Y_l - X_l Y_j).$$

PROOF. If Q can be written this way, then we will clearly have $Q(A(X, Y)) = (\det A)^2 Q(X, Y)$ for all A . Conversely, suppose this holds for all A , so that we also have

$$(2) \quad c_{ij,kl} = c_{kl,ij}$$

$$(3) \quad c_{li,jk} + c_{lj,ki} + c_{lk,ij} = 0.$$

We begin by writing four equivalent expressions for Q :

$$\begin{aligned} Q(X, Y) &= \sum c_{ij,kl} X_i X_j Y_k Y_l \\ &= \sum c_{jk,il} X_j X_k Y_i Y_l \\ &= \sum c_{il,jk} X_i X_l Y_j Y_k \\ &= \sum c_{kl,ij} X_k X_l Y_i Y_j. \end{aligned}$$

Now, by (3) we have

$$c_{jk,il} = -c_{ji,lk} - c_{jl,ki},$$

so

$$\begin{aligned} \sum_{i,j,k,l} c_{jk,il} X_j X_k Y_i Y_l &= - \sum_{i,j,k,l} c_{ji,lk} X_j X_k Y_i Y_l - \sum_{i,j,k,l} c_{jl,ki} X_j X_k Y_i Y_l \\ &= - \sum_{i,j,k,l} c_{ji,lk} X_j X_k Y_i Y_l - \sum_{i,j,k,l} c_{kl,ji} X_j X_k Y_i Y_l \\ &\quad \text{(interchanging } j \text{ and } k \text{ in the second sum)} \\ &= -2 \sum_{i,j,k,l} c_{ij,kl} X_j X_k Y_i Y_l, \quad \text{using (1) and (2).} \end{aligned}$$

If we apply a similar process to the third expression for Q , use (2) on the fourth, and leave the first unaltered, we obtain

$$\begin{aligned} Q(X, Y) &= \sum c_{ij,kl} X_i X_j Y_k Y_l \\ \frac{1}{2} Q(X, Y) &= - \sum c_{ij,kl} X_j X_k Y_i Y_l \\ \frac{1}{2} Q(X, Y) &= - \sum c_{ij,kl} X_i X_l Y_j Y_k \\ Q(X, Y) &= \sum c_{ij,kl} X_k X_l Y_i Y_j. \end{aligned}$$

Adding, we obtain the desired result,

$$3Q(X, Y) = \sum_{i,j,k,l} c_{ij,kl} (X_i Y_k - X_k Y_i) \cdot (X_j Y_l - X_l Y_j). \quad \blacklozenge$$

We now proceed with the hardest part of the investigation, a hairy calculation indeed.

4. PROPOSITION. In a Riemannian normal coordinate system x at p , the numbers

$$c_{ij,kl} = \frac{1}{2} \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}(p)$$

satisfy

$$\begin{aligned} c_{ij,kl} &= c_{kl,ij} \\ c_{li,jk} + c_{lj,ki} + c_{lk,ij} &= 0. \end{aligned}$$

PROOF. We begin with an equation derived in the proof of Proposition 1. For the geodesic $\gamma^k(t) = \xi^k t$ we have

$$\sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \xi^i \xi^j = 0;$$

multiplying by t^2 , we have

$$\sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) x^i(\gamma(t)) x^j(\gamma(t)) = 0.$$

Since these geodesics go through all points in a neighborhood of p , we have the following relation between the functions Γ_{ij}^k and x^i :

$$(1) \quad \sum_{i,j=1}^n \Gamma_{ij}^k x^i x^j = 0.$$

Since the tangent vector to the geodesic $\gamma^k(t) = \xi^k t$ has constant length, we also obtain

$$\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = \sum_{i=1}^n (\xi^i)^2,$$

which leads, in the same way, to the equation

$$(2) \quad \sum_{i,j=1}^n g_{ij} x^i x^j = \sum_{i=1}^n (x^i)^2.$$

Now equation (1) leads to

$$\sum_{i,j=1}^n [ij, k] x^i x^j = 0,$$

i.e., to

$$\sum_{i,j=1}^n \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right) x^i x^j = 0.$$

Interchanging the indices i and j in the second term, we can write

$$(3) \quad \sum_{i,j=1}^n \left(\frac{\partial g_{ik}}{\partial x^j} - \frac{1}{2} \frac{\partial g_{ij}}{\partial x^k} \right) x^i x^j = 0.$$

Our penultimate goal is to break this equation up into two sums, each of which is individually 0; the conditions on the c 's, which are our ultimate goal, will then follow fairly easily. To achieve this, our antepenultimate goal is to prove that $x^\beta = \sum_{\alpha} g_{\beta\alpha} x^\alpha$; these equations are at least reasonable, for they imply (2). To prove these relations, we first introduce the functions \bar{x}^β defined by

$$\bar{x}^\beta = \sum_{\alpha=1}^n g_{\beta\alpha} x^\alpha.$$

Note that

$$\frac{\partial \bar{x}^\beta}{\partial x^\gamma} = \sum_{\alpha=1}^n \frac{\partial g_{\beta\alpha}}{\partial x^\gamma} x^\alpha + g_{\beta\gamma}.$$

Substituting in (3), we obtain

$$\begin{aligned} 0 &= \sum_{j=1}^n \left(\sum_{i=1}^n \frac{\partial g_{ik}}{\partial x^j} x^i \right) x^j - \frac{1}{2} \sum_{i=1}^n \left(\sum_{j=1}^n \frac{\partial g_{ij}}{\partial x^k} x^j \right) x^i \\ &= \sum_{j=1}^n \left(\frac{\partial \bar{x}^k}{\partial x^j} - g_{kj} \right) x^j - \frac{1}{2} \sum_{i=1}^n \left(\frac{\partial \bar{x}^i}{\partial x^k} - g_{ik} \right) x^i \\ &= \sum_{j=1}^n \frac{\partial \bar{x}^k}{\partial x^j} x^j - \bar{x}^k - \frac{1}{2} \left(\sum_{i=1}^n \frac{\partial \bar{x}^i}{\partial x^k} x^i - \bar{x}^k \right) \\ &= \sum_{j=1}^n \frac{\partial \bar{x}^k}{\partial x^j} x^j - \frac{1}{2} \left(\sum_{i=1}^n \frac{\partial \bar{x}^i}{\partial x^k} x^i + \bar{x}^k \right) \\ &= \sum_{j=1}^n \frac{\partial \bar{x}^k}{\partial x^j} x^j - \frac{1}{2} \cdot \frac{\partial \left(\sum_{i=1}^n x^i \bar{x}^i \right)}{\partial x^k}. \end{aligned}$$

Now by (2) and the definition of \bar{x}^i , we have

$$\sum_{i=1}^n x^i \bar{x}^i = \sum_{i=1}^n (x^i)^2,$$

so we obtain

$$\begin{aligned} 0 &= \sum_{j=1}^n \frac{\partial \bar{x}^k}{\partial x^j} x^j - x^k \\ &= \sum_{j=1}^n \frac{\partial (\bar{x}^k - x^k)}{\partial x^j} x^j. \end{aligned}$$

This equation shows that along any geodesic $\gamma(t) = \xi^i t$ we have

$$\frac{d[\bar{x}^k - x^k](\gamma(t))}{dt} = 0,$$

so that $\bar{x}^k - x^k$ is constant along the geodesic. Since $g_{ij}(p) = \delta_{ij}$, we clearly have $\bar{x}^k(p) = x^k(p)$. Moreover, these geodesics pass through all points in a neighborhood of p . Thus $\bar{x}^k = x^k$ in a neighborhood of p , so that we finally obtain the desired equations

$$(4) \quad \sum_{\alpha=1}^n g_{k\alpha} x^\alpha = x^k.$$

Now we differentiate (4) to obtain

$$\sum_{\alpha=1}^n \frac{\partial g_{k\alpha}}{\partial x^l} x^\alpha + g_{kl} = \delta_{kl};$$

multiplying by x^l and summing, we obtain

$$\sum_{\alpha, l=1}^n \frac{\partial g_{k\alpha}}{\partial x^l} x^\alpha x^l = \sum_{l=1}^n -g_{kl} x^l + \delta_{kl} x^l,$$

which, together with (4) gives

$$\sum_{\alpha, l=1}^n \frac{\partial g_{k\alpha}}{\partial x^l} x^\alpha x^l = -x^k + x^k = 0,$$

and we have thus obtained the first part of our penultimate goal,

$$(5) \quad \sum_{i,j=1}^n \frac{\partial g_{ik}}{\partial x^j} x^i x^j = 0.$$

Together with (3), it implies the other part,

$$(6) \quad \sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^k} x^i x^j = 0.$$

We now obtain the desired equations as follows. Along the geodesic $\gamma^k(t) = \xi^k t$ we have, by (6),

$$(7) \quad \sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^k} (\gamma(t)) \xi^i \xi^j t^2 = 0.$$

This implies that

$$(8) \quad \sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^k} (\gamma(t)) \xi^i \xi^j = 0$$

for $t \neq 0$, and hence even for $t = 0$, by continuity. Differentiating (7) with respect to t gives

$$\begin{aligned} 0 &= \sum_{i,j=1}^n \frac{\partial g_{ij}}{\partial x^k} (\gamma(t)) \xi^i \xi^j \cdot 2t + \sum_{i,j,l=1}^n \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} (\gamma(t)) \xi^l \xi^i \xi^j t^2 \\ &= \sum_{i,j,l=1}^n \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} (\gamma(t)) \xi^l \xi^i \xi^j t^2, \quad \text{by (8);} \end{aligned}$$

consequently,

$$0 = \sum_{i,j,l=1}^n \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} (\gamma(t)) \xi^i \xi^j \xi^l$$

for all $t \neq 0$, and hence also for $t = 0$. Setting $t = 0$, we obtain

$$\sum_{i,j,l=1}^n \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} (p) \xi^i \xi^j \xi^l = 0.$$

This equation holds for *all* n -tuples ξ^1, \dots, ξ^n . From this we easily derive

$$(A) \quad c_{ij,kl} + c_{il,jk} + c_{jl,ik} = 0.$$

Applying the same process to (5), we obtain

$$(B) \quad c_{ki,jl} + c_{kj,li} + c_{kl,ij} = 0.$$

In (B) we interchange k and l , to obtain

$$c_{li,jk} + c_{lj,ki} + c_{kl,ij} = 0.$$

Comparing this equation with (A), we obtain the first of the desired relations,

$$c_{ij,kl} = c_{kl,ij}.$$

Moreover, using this relation with either (A) or (B), we obtain the second of the desired relations,

$$c_{li,jk} + c_{lj,ki} + c_{lk,ij} = 0.$$

And thus we are done! ♦

When we put all these results together we see that the quadratic function

$$Q(v_p, w_p) = \frac{1}{2} \sum_{i,j;k,l} \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} dx^i(v_p) dx^j(v_p) dx^k(w_p) dx^l(w_p)$$

can be written

$$Q(v_p, w_p) = \frac{1}{3} \sum_{i,j;k,l} c_{ij,kl} (dx^i \wedge dx^k) \cdot (dx^j \wedge dx^l)(v_p, w_p).$$

We thus see that the quadratic function Q , obtained from the Taylor expansion of $\| \cdot \|^2$ in Riemannian normal coordinates, has special properties which allow us to define, for any 2-dimensional subspace $W \subset M_p$, a number

$$Q(W) = \frac{Q(v_p, w_p)}{\|v_p, w_p\|^2} \quad v_p, w_p \text{ any basis for } W.$$

The work of the last four Propositions, which establishes this fact, is completely suppressed in Riemann's account, where the final result is merely stated, at the

beginning of section 2 of Part II. Riemann then makes some remarkable claims. First, Riemann interprets Q for a surface:

- (1) If M is 2-dimensional and $W = M_p$, then $-3Q(W)$ is just the Gaussian curvature $K(p)$ given by Theorem 3-7; we thus have an intrinsic definition of K , obtained by picking a special class of coordinate systems determined by the metric. (Riemann needs the factor $-3/4$ because he divides $Q(v_p, w_p)$ by the square of the area of the *triangle* spanned by v_p and w_p .)

At the end of section 2, Riemann interprets Q for an n -manifold:

- (2) If M is n -dimensional, $W \subset M_p$ is a 2-dimensional subspace, and $\mathcal{O} \subset W$ is a neighborhood of $0 \in W$ on which \exp is a diffeomorphism, then $-3Q(W)$ is the Gaussian curvature at p of the surface $\exp(\mathcal{O})$, with the metric it inherits as a submanifold of M .

But the most important claim is made in section 2. In an n -dimensional vector space there are $n\frac{n-1}{2}$ "independent" 2-dimensional subspaces: if v_1, \dots, v_n is a basis, we can choose the subspaces spanned by v_i and v_j , for $i < j$. Riemann claims that the metric $\langle \cdot, \cdot \rangle$ is determined if $Q(W)$ is known for $n\frac{n-1}{2}$ independent 2-dimensional subspaces $W \subset M_q$ at each point q , for example, if Q is known for the subspaces spanned by each $\partial/\partial x^i|_q$ and $\partial/\partial x^j|_q$ ($i < j$).

A very special case of this general claim is the following, which we will henceforth call the *Test Case*:

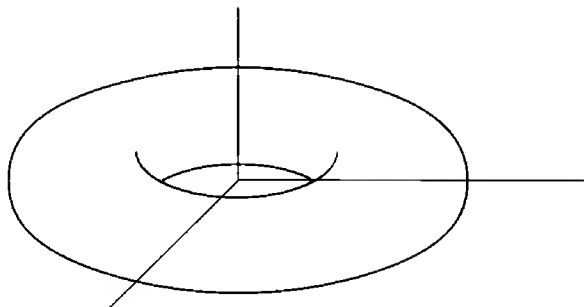
- (3) If M is n -dimensional and $Q = 0$ for $n\frac{n-1}{2}$ independent 2-dimensional subspaces of each M_q , then M is **flat**, that is, M is locally isometric to \mathbb{R}^n with its usual inner product.

In connection with the Test Case, it should be pointed out that a local isometry with \mathbb{R}^n is the best we can hope for, since there are Riemannian manifolds which are not homeomorphic to \mathbb{R}^n , but which are locally isometric to \mathbb{R}^n , and hence have $Q = 0$ everywhere. The simplest example of such a manifold, the "flat torus", is constructed as follows. The torus T can be obtained from \mathbb{R}^2 by identifying (x, y) with (x', y') if and only if

$$y' - y, x' - x \in \mathbb{Z}$$

(compare pg. I.372). The map $\pi: \mathbb{R}^2 \rightarrow T$, defined by taking (x, y) to its equivalence class, is locally a diffeomorphism, and there is clearly a unique metric $\langle \cdot, \cdot \rangle$ on T such that $\pi^*\langle \cdot, \cdot \rangle$ is the usual Riemannian metric on \mathbb{R}^2 ; consequently, $(T, \langle \cdot, \cdot \rangle)$ is locally isometric to \mathbb{R}^2 with its usual Riemannian

metric. Notice that the usual torus in \mathbb{R}^3 , with the induced Riemannian metric, is *not* flat; it has positive Gaussian curvature on the part furthest from the axis,



and negative Gaussian curvature on the part nearest the axis. However, if we consider $S^1 \subset \mathbb{R}^2$, then it is easy to see that $S^1 \times S^1 \subset \mathbb{R}^2 \times \mathbb{R}^2$, with the induced Riemannian metric, is flat.

One other remark should probably be made about the Test Case. At first sight, the Test Case might seem to be little more than a theorem about functions whose second partial derivatives are everywhere zero. However, it is actually quite different from this simple sort of result, since the value of Q at different points is defined in terms of different coordinate systems, each chosen specifically for one point.

Now our aim in the rest of this chapter is to prove assertions (1), (2), and (3). (The general claim that Q determines the metric will be considered later, as will the information given in sections 4 and 5 of Part II.) However, we will defer the proofs of assertions (1), (2), and (3) to another section of this chapter, not only in order to provide ourselves with a brief respite, but also to allow Riemann to add one or two more brilliant ideas.

C. A PRIZE ESSAY

The second edition of Riemann's collected works includes an unpublished paper, in Latin, which was submitted to the Paris Academy in 1861, to compete for a prize on a question involving heat conduction. In 1868, ten years after it had been offered, the prize was finally withdrawn. Because the way of obtaining the results of his essay were not fully explained, the prize was not awarded to Riemann, whose health prevented the more detailed handling of the subject which he had intended.

An extract from this paper is given below.* It should not be very hard to read, but the significance of the equations obtained there is only suggested by Riemann's final remarks; in the next part of this chapter we will have a great deal more to add. In the translation I have made some minor changes of notation.

An Extract From Riemann's Paper of 1861

Second Part

On the transformation of the expression $\sum_{i,j} g_{ij} dy^i dy^j$
into the given form $\sum_{i,j} a_{ij} dx^i dx^j$.

When the inquiry of the third Academy is restricted to homogeneous bodies, in which the resulting conductivities are constants, we develop the first condition that the expression $\sum_{i,j} g_{ij} dy^i dy^j$, in which the y^i are functions of the x^i , can be transformed into the form $\sum_{i,j} a_{ij} dx^i dx^j$ with given constant coefficients a_{ij} .

The expression $\sum_{i,j} a_{ij} dx^i dx^j$, if it is, as we shall suppose, a positive form in the dx^i , can always be put in the simplified form $\sum_i (dx^i)^2$. Thus if $\sum_{i,j} g_{ij} dy^i dy^j$ can be transformed into the form $\sum_{i,j} a_{ij} dx^i dx^j$, it can likewise be reduced to the form $\sum_i (dx^i)^2$ and vice versa. We therefore ask whether it can be put in the form $\sum_i (dx^i)^2$.

* Certain omissions, indicated by " . . . ", are considered in Addendum 2 to Chapter 6.

Let $G = \det(g_{ij})$ and let γ_{ij} be the cofactor; in this way $\sum_i g_{ij}\gamma_{ij} = G$ and $\sum_i g_{ij}\gamma_{ik} = 0$ if $j \neq k$.

If $\sum_{i,j} g_{ij} dy^i dy^j = \sum_i (dx^i)^2$ for arbitrary values of the dx^i , substituting $d + \delta$ for d leads also to $\sum_{i,j} g_{ij} dy^i \delta y^j = \sum_i dx^i \delta x^i$ for arbitrary values of the dx^i and δx^i .

Consequently, if the dy^i are expressed in terms of the dx^i and the δx^i in terms of the δy^i , it follows that

$$(1) \quad \frac{\partial x^\beta}{\partial y^\alpha} = \sum_i g_{\alpha i} \frac{\partial y^i}{\partial x^\beta}$$

and consequently

$$(2) \quad \frac{\partial y^i}{\partial x^\beta} = \sum_\alpha \frac{\gamma_{\alpha i}}{G} \frac{\partial x^\beta}{\partial y^\alpha}.$$

Thus we further deduce, seeing that

$$\sum_\alpha \frac{\partial y^i}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial y^i} = 1 \quad \text{and} \quad \sum_\alpha \frac{\partial y^i}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial y^j} = 0 \quad \text{if } i \neq j,$$

$$(3) \quad \sum_\alpha \frac{\partial x^\alpha}{\partial y^i} \frac{\partial x^\alpha}{\partial y^j} = g_{ij}, \quad (4) \quad \sum_\alpha \frac{\partial y^i}{\partial x^\alpha} \frac{\partial y^j}{\partial x^\alpha} = \frac{\gamma_{ij}}{G}$$

and differentiating formula (3),

$$\sum_\alpha \frac{\partial^2 x^\alpha}{\partial y^i \partial y^k} \frac{\partial x^\alpha}{\partial y^j} + \sum_\alpha \frac{\partial^2 x^\alpha}{\partial y^j \partial y^k} \frac{\partial x^\alpha}{\partial y^i} = \frac{\partial g_{ij}}{\partial y^k}.$$

Now from these expressions for

$$\frac{\partial g_{ij}}{\partial y^k}, \quad \frac{\partial g_{ik}}{\partial y^j}, \quad \frac{\partial g_{jk}}{\partial y^i}$$

we can write

$$(5) \quad 2 \sum_\alpha \frac{\partial^2 x^\alpha}{\partial y^j \partial y^k} \frac{\partial x^\alpha}{\partial y^i} = \frac{\partial g_{ij}}{\partial y^k} + \frac{\partial g_{ik}}{\partial y^j} - \frac{\partial g_{jk}}{\partial y^i}$$

and if these quantities are designated by p_{ijk} , then

$$(6) \quad 2 \frac{\partial^2 x^\alpha}{\partial y^j \partial y^k} = \sum_i \frac{\partial y^i}{\partial x^\alpha} p_{ijk}.$$

Differentiating the quantities p_{ijk} again yields

$$\frac{\partial p_{ijk}}{\partial y^l} - \frac{\partial p_{ijl}}{\partial y^k} = 2 \sum_\nu \frac{\partial^2 x^\nu}{\partial y^j \partial y^k} \frac{\partial^2 x^\nu}{\partial y^i \partial y^l} - 2 \sum_\nu \frac{\partial^2 x^\nu}{\partial y^j \partial y^l} \frac{\partial^2 x^\nu}{\partial y^i \partial y^k},$$

whence finally, substituting the values found in (6) and (4),

$$(I) \quad \frac{\partial^2 g_{ik}}{\partial y^j \partial y^l} + \frac{\partial^2 g_{jl}}{\partial y^i \partial y^k} - \frac{\partial^2 g_{il}}{\partial y^j \partial y^k} - \frac{\partial^2 g_{jk}}{\partial y^i \partial y^l} + \frac{1}{2} \sum_{\alpha, \beta} (p_{\alpha j l} p_{\beta i k} - p_{\alpha i l} p_{\beta j k}) \frac{\gamma_{\alpha \beta}}{G} = 0.$$

The functions g_{ij} must necessarily satisfy these equations whenever $\sum_{i,j} g_{ij} dy^i dy^j$ can be transformed into the form $\sum_i (dx^i)^2$: we denote the left side of this equation by

$$(ij, kl).$$

... Given an acquaintance with the traditional methods, it is demonstrated without difficulty that these ... conditions when they are satisfied, suffice

D. THE BIRTH OF THE RIEMANN CURVATURE TENSOR

All the developments in this part of the chapter have their origin in the following question, which Riemann considers in the paper of Part C: When is a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ *flat* (locally isometric to \mathbb{R}^n with its usual Riemannian metric)? In other words, when is there a coordinate system x^1, \dots, x^n on M for which

$$(1) \quad \langle \cdot, \cdot \rangle = \sum_{i=1}^n dx^i \otimes dx^i ?$$

We are going to seek an answer to this question in as straightforward a manner as possible; the quadratic function Q will not be used at all, but at the end it will make a surprise appearance.

We begin by choosing an arbitrary coordinate system y , in terms of which the metric $\langle \cdot, \cdot \rangle$ can be written

$$(2) \quad \langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} dy^i \otimes dy^j;$$

and we then seek conditions on the g_{ij} in order for (1) to hold for some coordinate system x . Since this is a purely local question, we can assume that y^1, \dots, y^n is just the standard coordinate system on \mathbb{R}^n .

If we express the dx^i in terms of the dy^j , and equate the coefficients of $dy^i \otimes dy^j$ in (2) with the resulting coefficients in (1), we find that the coordinate system x^1, \dots, x^n has the desired property if and only if

$$(3) \quad \sum_{\alpha} \frac{\partial x^{\alpha}}{\partial y^i} \frac{\partial x^{\alpha}}{\partial y^j} = g_{ij}.$$

From equation (3) we can immediately derive another, for we obtain

$$\begin{aligned} \sum_{j,\beta} g_{ij} \frac{\partial y^j}{\partial x^{\alpha}} \frac{\partial y^{\beta}}{\partial x^{\alpha}} &= \sum_{j,\beta,\alpha} \frac{\partial x^{\alpha}}{\partial y^i} \frac{\partial x^{\alpha}}{\partial y^j} \frac{\partial y^j}{\partial x^{\beta}} \frac{\partial y^{\beta}}{\partial x^{\alpha}} \\ &= \sum_{\alpha,\beta} \frac{\partial x^{\alpha}}{\partial y^i} \delta_{\beta}^{\alpha} \frac{\partial y^{\beta}}{\partial x^{\alpha}} \\ &= \sum_{\alpha} \frac{\partial x^{\alpha}}{\partial y^i} \frac{\partial y^{\alpha}}{\partial x^{\alpha}} = \delta_i^{\alpha}, \end{aligned}$$

which shows that if the coordinate system x^1, \dots, x^n has the desired property, then

$$(4) \quad \sum_{\beta} \frac{\partial y^i}{\partial x^{\beta}} \frac{\partial y^j}{\partial x^{\beta}} = g^{ij}.$$

Conversely, (4) implies (3). These results are just the equations (3) and (4) that Riemann obtains. Notice that Riemann begins with the square of the norm $\| \|^2 = \sum g_{ij} dy^i \cdot dy^j$, and then uses polarization to obtain the inner product, which he writes as $\sum g_{ij} dy^i \delta y^j$. Riemann also treats the two coordinate systems x and y on an equal footing throughout, so that his derivations of (3) and (4) are somewhat different. From (4) we obtain

$$\begin{aligned} \sum_{i,j} g^{ij} \frac{\partial x^{\mu}}{\partial y^i} \frac{\partial x^{\nu}}{\partial y^j} &= \sum_{\beta,i,j} \frac{\partial y^i}{\partial x^{\beta}} \frac{\partial y^j}{\partial x^{\beta}} \frac{\partial x^{\mu}}{\partial y^i} \frac{\partial x^{\nu}}{\partial y^j} \\ &= \sum_{\beta} \delta_{\beta}^{\mu} \delta_{\beta}^{\nu}, \end{aligned}$$

and thus the coordinate system x^1, \dots, x^n has the desired property if and only if

$$(4') \quad \sum_{i,j} g^{ij} \frac{\partial x^{\mu}}{\partial y^i} \frac{\partial x^{\nu}}{\partial y^j} = \delta_{\mu\nu}.$$

This equation, which we will find more useful than (4), can be derived directly from (3) in the following way. If $A = (a_{ij}) = (\partial x^i / \partial y^j)$, and $G = (g_{ij})$, then (3) says that

$$A^t \cdot A = G,$$

where A^t is the transpose of A ; this is equivalent to

$$G^{-1} = A^{-1} \cdot (A^t)^{-1},$$

and hence to

$$AG^{-1}A^t = I,$$

which is just (4'). In particular, this shows immediately that (4') is equivalent to (3).

Now equation (3) is a partial differential equation for the functions x^{α} . In Chapter I.6 we developed a general theory for partial differential equations, but we notice at once that (3) is not an equation of the type to which our theory

applies. Our first task will thus be to obtain from (3) an equation that we do know how to handle. The situation is very much like, and may profitably be compared to, that which occurs in Problem I.7-19, where the analysis of a certain set of partial differential equations is reduced to Theorem I.6-1, together with the Poincaré Lemma. To treat equation (3), we begin by differentiating (about all we can do), to obtain

$$\sum_{\alpha} \frac{\partial^2 x^{\alpha}}{\partial y^i \partial y^k} \frac{\partial x^{\alpha}}{\partial y^j} + \sum_{\alpha} \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} \frac{\partial x^{\alpha}}{\partial y^i} = \frac{\partial g_{ij}}{\partial y^k}.$$

By writing down this equation for

$$\frac{\partial g_{ij}}{\partial y^k}, \quad \frac{\partial g_{ik}}{\partial y^j}, \quad \frac{\partial g_{jk}}{\partial y^i},$$

and combining, we obtain an equation equivalent to Riemann's,

$$(5) \quad \sum_{\alpha} \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} \frac{\partial x^{\alpha}}{\partial y^i} = \frac{1}{2} \left(\frac{\partial g_{ij}}{\partial y^k} + \frac{\partial g_{ik}}{\partial y^j} - \frac{\partial g_{jk}}{\partial y^i} \right) = [jk, i].$$

Thus, the symbols $[jk, i]$, which came up naturally in the calculus of variations, also come up naturally in this different context. After Riemann's *Habilitations* lecture was published, in 1866, several mathematicians independently derived his results or considered related questions. Christoffel, in particular, introduced these combinations of the partial derivatives of the g_{ij} 's, and the symbols $[ij, k]$ and Γ_{ij}^k are called the **Christoffel symbols** of the **first** and **second kinds**, respectively (Christoffel actually used $[ij]_k$ and $\{ij\}^k$, which do not accommodate themselves to the summation convention). In the next chapter we will see one important use which Christoffel made of these symbols.

At this point we will depart slightly from Riemann's treatment, in order to obtain equations to which Theorem I.6-1 directly applies. From (5) we obtain

$$\begin{aligned} \sum_{i, \gamma} g^{i\gamma} \frac{\partial x^{\lambda}}{\partial y^{\gamma}} [jk, i] &= \sum_{\alpha, i, \gamma} \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} \frac{\partial x^{\alpha}}{\partial y^i} \frac{\partial x^{\lambda}}{\partial y^{\gamma}} g^{i\gamma} \\ &= \sum_{\alpha} \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} \left(\sum_{i, \gamma} \frac{\partial x^{\alpha}}{\partial y^i} \frac{\partial x^{\lambda}}{\partial y^{\gamma}} g^{i\gamma} \right) \\ &= \sum_{\alpha} \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} \delta_{\alpha\lambda} \quad \text{by (4')}, \end{aligned}$$

so we obtain, finally,

$$(6) \quad \frac{\partial^2 x^\lambda}{\partial y^j \partial y^k} = \sum_{\gamma=1}^n \Gamma_{jk}^\gamma \frac{\partial x^\lambda}{\partial y^\gamma}$$

(which is easily seen to be equivalent to Riemann's equation (6)); we will also write this equation as

$$\frac{\partial \left(\frac{\partial x^\lambda}{\partial y^j} \right)}{\partial y^k} = \sum_{\gamma=1}^n \Gamma_{jk}^\gamma \frac{\partial x^\lambda}{\partial y^\gamma}.$$

Notice that the index λ plays no special role here; all functions x^λ satisfy the *same* equation. Thus, for each λ the n -tuple of functions

$$\alpha = \left(\frac{\partial x^\lambda}{\partial y^1}, \dots, \frac{\partial x^\lambda}{\partial y^n} \right) \quad \alpha: \mathbb{R}^n \rightarrow \mathbb{R}^n$$

satisfies the set of partial differential equations

$$(*) \quad \frac{\partial \alpha}{\partial y^k}(y) = f_k(y, \alpha(y)),$$

where $f_k: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is given by

$$f_k^j(y, z) = \sum_{\gamma=1}^n \Gamma_{jk}^\gamma(y) \cdot z^\gamma.$$

Since this is true for every λ , the equation $(*)$ has n solutions whose initial values at some point, 0 say, are linearly independent. Since constant linear combinations of solutions of $(*)$ are also solutions, it follows that $(*)$ has solutions with arbitrary initial conditions at 0. From Theorem I.6-1 we thus obtain necessary integrability conditions,

$$\frac{\partial f_k}{\partial y^l} - \frac{\partial f_l}{\partial y^k} + \sum_{\mu=1}^n \frac{\partial f_k}{\partial z^\mu} f_l^\mu - \sum_{\mu=1}^n \frac{\partial f_l}{\partial z^\mu} f_k^\mu = 0.$$

In our case, looking at the j^{th} components of these equations, we obtain

$$\sum_{\gamma=1}^n \frac{\partial \Gamma_{jk}^\gamma}{\partial y^l} z^\gamma - \sum_{\gamma=1}^n \frac{\partial \Gamma_{jl}^\gamma}{\partial y^k} z^\gamma + \sum_{\mu=1}^n \Gamma_{jk}^\mu \sum_{\gamma=1}^n \Gamma_{\mu l}^\gamma z^\gamma - \sum_{\mu=1}^n \Gamma_{jl}^\mu \sum_{\gamma=1}^n \Gamma_{\mu k}^\gamma z^\gamma = 0.$$

Since these relations must hold for all $z = (z^1, \dots, z^n)$, we obtain

$$(**) \quad 0 = R^\gamma_{jlk} \stackrel{\text{def}}{=} \frac{\partial \Gamma^\gamma_{kj}}{\partial y^l} - \frac{\partial \Gamma^\gamma_{lj}}{\partial y^k} + \sum_{\mu=1}^n (\Gamma^\mu_{kj} \Gamma^\gamma_{l\mu} - \Gamma^\mu_{lj} \Gamma^\gamma_{k\mu})$$

as necessary conditions that $\sum g_{ij} dy^i \otimes dy^j = \sum dx^i \otimes dx^i$ for some coordinate system $x = (x^1, \dots, x^n)$. Notice that the set of equations $R^\gamma_{jlk} = 0$ is equivalent to the set of equations

$$R_{ijlk} \stackrel{\text{def}}{=} \sum_{\gamma=1}^n g_{i\gamma} R^\gamma_{jlk} = 0.$$

The quantities R_{ijlk} can be expressed in another way, after a little calculation. Note first that

$$\begin{aligned} \sum_{\gamma=1}^n g_{i\gamma} \frac{\partial \Gamma^\gamma_{jk}}{\partial y^l} &= \frac{\partial}{\partial y^l} \left(\sum_{\gamma=1}^n g_{i\gamma} \Gamma^\gamma_{jk} \right) - \sum_{\gamma=1}^n \Gamma^\gamma_{jk} \frac{\partial g_{i\gamma}}{\partial y^l} \\ &= \frac{\partial [jk, i]}{\partial y^l} - \sum_{\gamma=1}^n \Gamma^\gamma_{jk} ([il, \gamma] + [\gamma l, i]). \end{aligned}$$

Substituting into (**), and remembering the definition of $[ij, k]$, we obtain

$$(***) \quad \begin{aligned} R_{ijlk} &= \frac{1}{2} \left(\frac{\partial^2 g_{ik}}{\partial y^j \partial y^l} + \frac{\partial^2 g_{jl}}{\partial y^i \partial y^k} - \frac{\partial^2 g_{il}}{\partial y^j \partial y^k} - \frac{\partial^2 g_{jk}}{\partial y^i \partial y^l} \right) \\ &\quad + \sum_{\alpha, \beta=1}^n g^{\alpha\beta} ([jl, \alpha] \cdot [ik, \beta] - [il, \alpha] \cdot [jk, \beta]). \end{aligned}$$

The condition $R_{ijlk} = 0$ is just the condition (I) which Riemann obtains (note that Riemann's p_{ijk} equals $2[jk, i]$)—the quantity which we have denoted by R_{ijlk} is what Riemann denotes by $2(ij, kl)$; the factor of 2 is not particularly significant, nor is the interchange of l and k , for it is easily seen that $R_{ijlk} = -R_{ijkl}$.

The notation R^i_{jkl} has been picked in anticipation of the following result.

5. PROPOSITION. On a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ there is a tensor of type $\binom{3}{1}$ whose components in any coordinate system y are

$$R^i_{jkl} = \frac{\partial \Gamma^i_{jl}}{\partial y^k} - \frac{\partial \Gamma^i_{jk}}{\partial y^l} + \sum_{\mu=1}^n (\Gamma^\mu_{jl} \Gamma^i_{\mu k} - \Gamma^\mu_{jk} \Gamma^i_{\mu l})$$

(where $\langle \cdot, \cdot \rangle = \sum g_{ij} dy^i \otimes dy^j$, and the Christoffel symbols Γ are defined as usual).

PROOF. We just compute that the components transform “correctly”!! In other words, if $R'^i{}_{jlk}$ are defined by the same formula, with respect to the coordinate system y' , we show that

$$R'^{\alpha}{}_{\beta\gamma\delta} = \sum_{i,j,k,l} R^i{}_{jkl} \frac{\partial y^j}{\partial y'^{\beta}} \frac{\partial y^k}{\partial y'^{\gamma}} \frac{\partial y^l}{\partial y'^{\delta}} \frac{\partial y'^{\alpha}}{\partial y^i}.$$

To do this, all one needs is the result from Problem I.9-22,

$$\Gamma'^{\nu}{}_{\alpha\beta} = \sum_{i,j,k} \Gamma^k{}_{ij} \frac{\partial y^i}{\partial y'^{\alpha}} \frac{\partial y^j}{\partial y'^{\beta}} \frac{\partial y'^{\nu}}{\partial y^k} + \sum_{\mu=1}^n \frac{\partial^2 y^{\mu}}{\partial y'^{\alpha} \partial y'^{\beta}} \frac{\partial y'^{\nu}}{\partial y^{\mu}},$$

and plenty of perseverance.

SLIGHTLY MORE MOTIVATED PROOF. Begin with the equation

$$\sum_{i,j} g_{ij} dy^i \otimes dy^j = \langle \ , \ \rangle = \sum_{i,j} g'_{ij} dy'^i \otimes dy'^j,$$

and repeat the whole sequence of computations which we performed in the special case that $g'_{ij} = \delta_{ij}$. The result will be the desired transformation law. (The integrability conditions (**)) then follow as a necessary condition for the existence of a coordinate system y' with $g'_{ij} = \delta_{ij}$, for in such a coordinate system we clearly have $R'^i{}_{jkl} = 0$, which in turn implies that all $R^i{}_{jkl} = 0$.) ♦

We have thus stumbled onto a new tensor, the **Riemann curvature tensor**, which in the coordinate system y equals

$$\sum_{i,j,k,l} R^i{}_{jkl} dy^j \otimes dy^k \otimes dy^l \otimes \frac{\partial}{\partial y^i}.$$

Eventually we hope to have a useful invariant definition of this tensor; this will involve an enormous amount of exploration. For the time being, we simply accept the classical definition, which arises naturally as an integrability condition, and explain how it is connected with curvature. In the process we will obtain an invariant, but extraordinarily clumsy, definition of the curvature tensor.

It will be convenient to introduce a bit of modern terminology, and denote by R the tensor with components $R^i{}_{jkl}$. Since this tensor is of type $\binom{3}{1}$ it may be regarded as a function taking three vectors to another vector. The value of R on $X, Y, Z \in M_p$ will be denoted by

$$R(Y, Z)X \in M_p,$$

and hence we have

$$R\left(\left.\frac{\partial}{\partial y^k}\right|_p, \left.\frac{\partial}{\partial y^l}\right|_p\right)\left.\frac{\partial}{\partial y^j}\right|_p = \sum_{i=1}^n R^i{}_{jkl}(p) \cdot \left.\frac{\partial}{\partial y^i}\right|_p$$

(the reason for choosing the notation $R(Y, Z)X$ comes out in Proposition 6).

The numbers $R_{ijkl} = \sum_{\gamma} g_{i\gamma} R^{\gamma}{}_{jkl}$ are also the components of a tensor, of type $\binom{4}{0}$, but it is unnecessary to perform any calculations to verify this. Clearly

$$R_{ijkl}(p) = \left\langle R\left(\left.\frac{\partial}{\partial y^k}\right|_p, \left.\frac{\partial}{\partial y^l}\right|_p\right)\left.\frac{\partial}{\partial y^j}\right|_p, \left.\frac{\partial}{\partial y^i}\right|_p \right\rangle,$$

so the tensor in question is just the multilinear map

$$(X, Y, Z, W) \mapsto \langle R(Z, W)Y, X \rangle.$$

This function of four tangent vectors is closely connected with the quadratic function introduced in Part B of this chapter:

6. PROPOSITION. Let x be a Riemannian normal coordinate system at p , and Q the quadratic function on $M_p \times M_p$ defined by

$$Q(X, Y) = \sum_{i,j,k,l} c_{ij,kl} dx^i(X) dx^j(X) dx^k(Y) dx^l(Y),$$

where

$$c_{ij,kl} = \frac{1}{2} \frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}.$$

Then

$$Q(X, Y) = -\frac{1}{3} \langle R(X, Y)Y, X \rangle.$$

PROOF. We have seen that

$$\begin{aligned} 3Q(X, Y) &= \sum_{i,j,k,l} c_{ij,kl} (dx^i \wedge dx^k) \cdot (dx^j \wedge dx^l)(X, Y) \\ &= \sum_{i,j,k,l} c_{ij,kl} dx^i(X) dx^j(X) dx^k(Y) dx^l(Y) \\ &\quad + \sum_{i,j,k,l} c_{ij,kl} dx^k(X) dx^l(X) dx^i(Y) dx^j(Y) \\ &\quad - \sum_{i,j,k,l} c_{ij,kl} dx^j(X) dx^k(X) dx^i(Y) dx^l(Y) \\ &\quad - \sum_{i,j,k,l} c_{ij,kl} dx^i(X) dx^l(X) dx^j(Y) dx^k(Y). \end{aligned}$$

By switching indices we can rewrite this as

$$\begin{aligned}
 3Q(X, Y) &= \sum_{i,j,k,l} c_{ik,jl} dx^i(X) dx^j(Y) dx^k(X) dx^l(Y) && \text{[interchange } j \text{ and } k] \\
 &+ \sum_{i,j,k,l} c_{jl,ik} && \text{[interchange } i \text{ and } l] \\
 &- \sum_{i,j,k,l} c_{il,jk} && \text{[change } i \text{ to } l; j \text{ to } i; l \text{ to } j] \\
 &- \sum_{i,j,k,l} c_{jk,il} && \text{[change } i \text{ to } k; k \text{ to } l; l \text{ to } i] \\
 &= \sum_{i,j,k,l} (c_{ik,jl} + c_{jl,ik} - c_{il,jk} - c_{jk,il}) dx^i \otimes dx^j \otimes dx^k \otimes dx^l(X, Y, X, Y).
 \end{aligned}$$

Now in Riemannian normal coordinates, the Christoffel symbols $[ij, k]$ are all 0 at p , since all $\partial g_{ij}/\partial x^k$ are 0 at p . Referring to equation (***) we thus have

$$\begin{aligned}
 3Q(X, Y) &= \sum_{i,j,k,l} R_{ijkl}(p) dx^i \otimes dx^j \otimes dx^k \otimes dx^l(X, Y, X, Y) \\
 &= - \sum_{i,j,k,l} R_{ijkl}(p) dx^i \otimes dx^j \otimes dx^k \otimes dx^l(X, Y, X, Y) \\
 &= -\langle R(X, Y)Y, X \rangle. \spadesuit
 \end{aligned}$$

We are now ready to verify some of Riemann's claims.

7. PROPOSITION. Let $(M, \langle \cdot, \cdot \rangle)$ be a 2-dimensional Riemannian manifold, and let $X, Y \in M_p$ be linearly independent. Let $\|X, Y\|$ denote the area of the parallelogram spanned by X and Y . Then

$$K(p) = \frac{\langle R(X, Y)Y, X \rangle}{\|X, Y\|^2} \quad [= \langle R(X, Y)Y, X \rangle \text{ if } X \text{ and } Y \text{ are orthonormal}]$$

is the same as the Gaussian curvature at p defined by the formula in Theorem 3-7 (in particular, this proves that the formula in Theorem 3-7 is indeed independent of the coordinate system).

FIRST PROOF. Let (x, y) be a coordinate system on a neighborhood of p . It obviously suffices to verify the theorem when $X = \partial/\partial x|_p$ and $Y = \partial/\partial y|_p$,

since by Proposition 6, and the results of Part B, the numerator is multiplied by the same factor as the denominator when we change to any other pair of vectors. In this case,

$$\begin{aligned}\langle R(X, Y)Y, X \rangle &= \left\langle R\left(\frac{\partial}{\partial x}\Big|_p, \frac{\partial}{\partial y}\Big|_p\right) \frac{\partial}{\partial y}\Big|_p, \frac{\partial}{\partial x}\Big|_p \right\rangle \\ &= R_{1212}(p).\end{aligned}$$

If we write

$$\langle \cdot, \cdot \rangle = E dx \otimes dx + F dx \otimes dy + F dy \otimes dx + G dy \otimes dy,$$

so that

$$\begin{aligned}g_{11} &= E \\ g_{12} &= g_{21} = F \\ g_{22} &= G,\end{aligned}$$

then (by the formula on pg. I.308)

$$\|X, Y\|^2 = EG - F^2,$$

so we must prove that

$$4R_{1212}(EG - F^2) = 4(EG - F^2)^2 K,$$

where the right side is given by the formula in Theorem 3-7. This is a fairly straightforward calculation from (***) on page 188. The first term in (***) corresponds to the last in the formula for $4(EG - F^2)^2 K$, and the second corresponds to the first three in the latter formula. In carrying out the calculation, note that

$$\begin{aligned}g^{11} &= \frac{G}{EG - F^2} \\ g^{12} &= g^{21} = \frac{-F}{EG - F^2} \\ g^{22} &= \frac{E}{EG - F^2};\end{aligned}$$

the denominators cancel out the unwanted factor in $4R_{1212}(EG - F^2)$.

SECOND PROOF (OUTLINE). Let (r, φ) be the coordinate system around p which is introduced on page 136. We know that in this coordinate system

$$\langle \cdot, \cdot \rangle = dr \otimes dr + G d\varphi \otimes d\varphi$$

for some function G , and (see page 145) that

$$K(p) = -\frac{\partial^3 \sqrt{G}}{\partial r^3}(p).$$

Introduce a Riemannian normal coordinate system x^1, x^2 by the equations

$$x^1 = r \cos \varphi, \quad x^2 = r \sin \varphi.$$

We can then calculate the g_{ij} in terms of G , and use these results to show that the quantity

$$\mathcal{Q} \left(\left. \frac{\partial}{\partial x^1} \right|_p, \left. \frac{\partial}{\partial x^2} \right|_p \right) = 2c_{11,22}(p)$$

is equal to $-K(p)/3$. The result then follows from Proposition 6. ♦

8. PROPOSITION. Let $(M, \langle \cdot, \cdot \rangle)$ be a Riemannian manifold, and let W be a 2-dimensional subspace of M_p , spanned by $X, Y \in M_p$. Let $\mathcal{O} \subset W$ be a neighborhood of $0 \in M_p$ on which \exp is a diffeomorphism, let $i: \exp(\mathcal{O}) \rightarrow M$ be the inclusion, and let \bar{R} be the Riemann curvature tensor for $\exp(\mathcal{O})$ with the induced Riemannian metric $i^*\langle \cdot, \cdot \rangle$. Then

$$\langle \bar{R}(X, Y)Y, X \rangle = \langle R(X, Y)Y, X \rangle.$$

Consequently,

$$\frac{\langle R(X, Y)Y, X \rangle}{\|X, Y\|^2}$$

is the Gaussian curvature at p of the surface $\exp(\mathcal{O})$.

FIRST PROOF. It obviously suffices to prove the theorem when X and Y are orthonormal. Choose a Riemannian normal coordinate system at p with $X = \partial/\partial x^1|_p$, $Y = \partial/\partial x^2|_p$; then x^1, x^2 is a coordinate system on $\exp(\mathcal{O})$. Now we are trying to prove that $\bar{R}_{1212}(p) = R_{1212}(p)$. But in (**), the terms involving Christoffel symbols vanish at p . The theorem is now obvious, since the functions \bar{g}_{ij} ($i, j = 1, 2$) defining the metric $i^*\langle \cdot, \cdot \rangle$ are just the corresponding g_{ij} restricted to $\exp(\mathcal{O})$, and they have the same mixed partial derivatives with respect to x^1 and x^2 .

SECOND PROOF. It is even more obvious that the quadratic form \bar{Q} associated with $(\exp(\mathcal{O}), i^*\langle \cdot, \cdot \rangle)$ is the restriction to W of the quadratic form Q on M_p , for they are the second non-zero terms in the Taylor expansion of the same metric. ♦

9. COROLLARY. Let $(M, \langle \cdot, \cdot \rangle)$ be a Riemannian manifold, let $X, Y \in M_p$ span a 2-dimensional subspace W of M_p , and let $\mathcal{O} \subset W$ be a neighborhood of 0 on which \exp is a diffeomorphism. If Q is the quadratic form on M defined previously, then

$$\frac{-3Q(X, Y)}{\|X, Y\|^2} = \frac{\langle R(X, Y)Y, X \rangle}{\|X, Y\|^2} = K,$$

where K is the Gaussian curvature at p of the surface $\exp(\mathcal{O})$.

The quantity $\langle R(X, Y)Y, X \rangle / \|X, Y\|^2$ appearing in Corollary 9 is called the **sectional curvature** $K(W)$ of W . It would seem that the function $(X, Y) \mapsto \langle R(X, Y)Y, X \rangle$ contains only a small portion of the total information contained in the curvature tensor, but Propositions 10 and 12, which follow, show that R satisfies certain identities which allow it to be determined in terms of the metric $\langle \cdot, \cdot \rangle$ and the quadratic function Q which it determines.

10. PROPOSITION. The curvature tensor satisfies the following identities:

(1) $R(X, Y)Z = -R(Y, X)Z$, hence

$$\langle R(X, Y)Z, W \rangle = -\langle R(Y, X)Z, W \rangle.$$

(2) $\langle R(X, Y)Z, W \rangle = -\langle R(X, Y)W, Z \rangle$.

(3) $R(X, Y)Z + R(Y, Z)X + R(Z, X)Y = 0$, hence

$$\langle R(X, Y)Z, W \rangle + \langle R(Y, Z)X, W \rangle + \langle R(Z, X)Y, W \rangle = 0.$$

(4) $\langle R(X, Y)Z, W \rangle = \langle R(Z, W)X, Y \rangle$.

PROOF. In a coordinate system x , these relations are equivalent to

$$(1) \quad R^i_{jkl} = -R^i_{jlk} \quad \text{or} \quad R_{ijkl} = -R_{ijlk}$$

$$(2) \quad R_{ijkl} = -R_{jikl}$$

$$(3) \quad R^i_{jkl} + R^i_{klj} + R^i_{ljk} = 0 \quad \text{or} \quad R_{ijkl} + R_{iklj} + R_{iljk} = 0$$

$$(4) \quad R_{klij} = R_{ijkl}.$$

These are immediate from (**) and (***). ♦

Notice that $\langle R(X, Y)Z, W \rangle$ is skew-symmetric in both (X, Y) and (Z, W) , which again shows that $\langle R(X, Y)Y, X \rangle$ changes by $\det(a_{ij})^2$ when X and Y are replaced by $a_{11}X + a_{21}Y, a_{21}X + a_{22}Y$. For later use, we insert a result which shows that the fourth property of R is a formal consequence of the others.

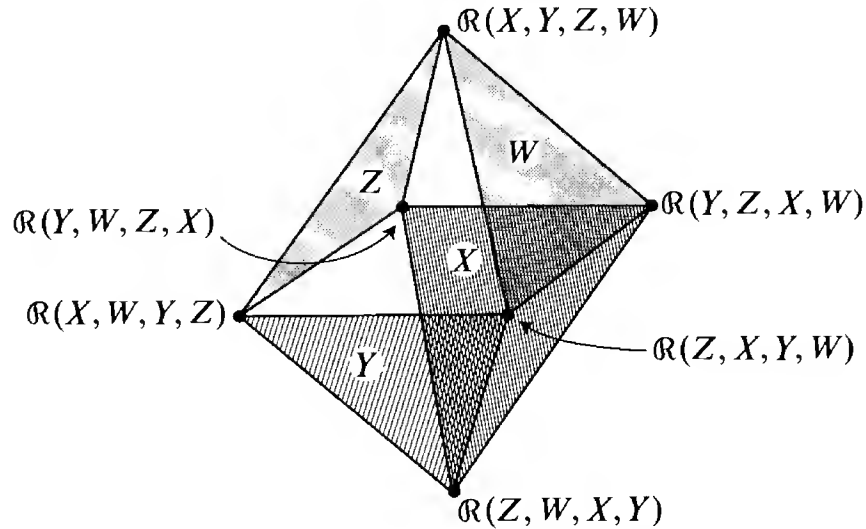
11. PROPOSITION. Let V be a vector space and $\mathcal{R}: V \times V \times V \times V \rightarrow \mathbb{R}$ a multilinear map satisfying

- (1) $\mathcal{R}(X, Y, Z, W) = \mathcal{R}(Y, X, Z, W)$
- (2) $\mathcal{R}(X, Y, Z, W) = \mathcal{R}(X, Y, W, Z)$
- (3) $\mathcal{R}(X, Y, Z, W) + \mathcal{R}(Y, Z, X, W) + \mathcal{R}(Z, X, Y, W) = 0.$

Then \mathcal{R} also satisfies

- (4) $\mathcal{R}(X, Y, Z, W) = \mathcal{R}(Z, W, X, Y).$

PROOF. The proof is a tricky manipulation, cleverly systematized by the following diagram from Milnor's *Morse Theory*.



Equation (3) shows that the sum of the numbers at the vertices of triangle W is zero. The sums of the vertices of triangles X , Y , and Z are also seen to be zero, using (1) and (2). Adding these identities for the top two triangles, and subtracting the identities for the bottom ones, we see that twice the top vertex minus twice the bottom vertex is zero. ♦

12. PROPOSITION. Let V be a vector space and $\mathcal{R}_i: V \times V \times V \times V \rightarrow \mathbb{R}$ two multilinear maps satisfying (1)–(4) of Proposition 11. Suppose $\mathcal{R}_1(X, Y, X, Y) = \mathcal{R}_2(X, Y, X, Y)$ for all $X, Y \in V$. Then $\mathcal{R}_1 = \mathcal{R}_2$.

PROOF. It clearly suffices to prove that a multilinear \mathcal{R} satisfying (1)–(4) is 0 if $\mathcal{R}(X, Y, X, Y) = 0$ for all $X, Y \in V$. Now we have

$$\begin{aligned}
 0 &= \mathcal{R}(X, Y + W, X, Y + W) \\
 &= \mathcal{R}(X, Y, X, Y) + \mathcal{R}(X, Y, X, W) + \mathcal{R}(X, W, X, Y) + \mathcal{R}(X, W, X, W) \\
 &= \mathcal{R}(X, Y, X, W) + \mathcal{R}(X, W, X, Y) \\
 &= 2\mathcal{R}(X, Y, X, W).
 \end{aligned}$$

Using (1) and (2), we easily see that \mathcal{R} is alternating, and hence skew-symmetric. Consequently, (3) gives

$$3\mathcal{R}(X, Y, Z, W) = 0. \spadesuit$$

Propositions 10 and 12 tell us that the curvature tensor R is completely determined by the values of $\langle R(X, Y)Y, X \rangle$, and hence by the quadratic function Q . [This means that in a sense we can frame a coordinate-free definition of the curvature tensor, but it would certainly be an awkward one. Moreover, given a multilinear map $\mathcal{R}: V \times V \times V \times V \rightarrow \mathbb{R}$, satisfying (1)–(4), it is a fairly difficult exercise to work out a formula for \mathcal{R} in terms of the quantities $\mathcal{R}(X, Y, X, Y)$.] In terms of a coordinate system, we see that the tensor $\mathcal{R}(X, Y, Z, W) = \langle R(X, Y)Z, W \rangle$ is determined by the components R_{ijij} , of which there are $n\frac{n-1}{2}$ with $i < j$. According to Riemann, these $n\frac{n-1}{2}$ functions must determine the metric completely; in other words, the tensor \mathcal{R} must determine the metric.*

Recall that we have selected one special case of this assertion as our Test Case, which can now be restated as follows: If $R = 0$, then the manifold is flat. We are ready to present the first, and longest, of our proofs of the Test Case. It is separated into three Steps, and all our subsequent proofs, no matter how elegant and brief, essentially contain these same three Steps.

Recall that for a coordinate system y^1, \dots, y^n we have the formula (pg. I.331)

$$(*) \quad \frac{\partial g_{ij}}{\partial y^k} = [ik, j] + [jk, i],$$

which is equivalent to the definition of the Christoffel symbols, as well as the formula (pg. I.331)

$$(**) \quad \frac{\partial g^{ij}}{\partial y^k} = - \sum_{l=1}^n (g^{il} \Gamma_{lk}^j + g^{lj} \Gamma_{lk}^i),$$

which can be derived from it.

*This is not really the same as saying that R determines the metric, since we can't determine $R(X, Y)Z$ from $\mathcal{R}(X, Y, Z, W) = \langle R(X, Y)Z, W \rangle$ unless the metric is already known! In fact, any numbers R_{ijkl} satisfying the identities of Proposition 6 can be realized as the components, at a point, of R for some metric (the R_{ijkl} determine the second derivatives of the metric at the point).

13. THEOREM (THE TEST CASE; FIRST VERSION). Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. This is a purely local question, so we assume that M is \mathbb{R}^n , with the standard coordinate system y^1, \dots, y^n , and the Riemannian metric

$$\langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} dy^i \otimes dy^j.$$

Step 1. We claim that there are functions (h_1, \dots, h_n) , with any desired initial conditions $(h_1(0), \dots, h_n(0))$, satisfying the equations

$$(*) \quad \frac{\partial h_j}{\partial y^k} = \sum_{\gamma=1}^n \Gamma_{jk}^{\gamma} h_{\gamma}.$$

The reason for this is, of course, that the relations $R^{\gamma}_{jlk} = 0$, which express the vanishing of R , are just the integrability conditions for $(*)$, as we have already seen.

In particular, for $\alpha = 1, \dots, n$ we can choose such a set $(h^{(\alpha)}_1, \dots, h^{(\alpha)}_n)$ satisfying the initial condition

$$(h^{(\alpha)}_1(0), \dots, h^{(\alpha)}_n(0))_0 = X_{\alpha},$$

where $X_1, \dots, X_n \in \mathbb{R}^n_0$ is orthonormal with respect to $\langle \cdot, \cdot \rangle_0$.

Step 2. We claim that if (h_1, \dots, h_n) satisfies $(*)$, then $h = dx$ for some function x , i.e., $h_j = \partial x / \partial y^j$. In terms of the form

$$\eta = h_1 dy^1 + \dots + h_n dy^n,$$

we are just saying that η is exact. We know (Corollary I.7-15) that this is true if and only if

$$\frac{\partial h_j}{\partial y^k} = \frac{\partial h_k}{\partial y^j}.$$

Glancing at $(*)$, we see that this is indeed true, since $\Gamma_{jk}^{\gamma} = \Gamma_{kj}^{\gamma}$.

Now choose functions x^{α} with $h^{(\alpha)}_j = \partial x^{\alpha} / \partial y^j$. Then the functions x^{α} satisfy

$$(\dagger) \quad \frac{\partial^2 x^{\alpha}}{\partial y^j \partial y^k} = \sum_{\gamma=1}^n \Gamma_{jk}^{\gamma} \frac{\partial x^{\alpha}}{\partial y^{\gamma}} \quad \left[\begin{array}{l} \text{these are the equations (6),} \\ \text{obtained earlier, page 187} \end{array} \right]$$

and

$$\left(\frac{\partial x^\alpha}{\partial y^j}(0) \right) = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix};$$

this matrix is non-singular, so x^1, \dots, x^n is a coordinate system in a neighborhood of 0.

Step 3. We claim that x is the desired coordinate system, i.e., that

$$(\dagger) \quad \delta_{\mu\nu} = \sum_{i,j=1}^n g^{ij} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} \quad [\text{equation (4'), page 185}].$$

We know that this equation holds at 0, by the choice of the initial conditions $\partial x^\alpha / \partial y^j(0)$. So it suffices to show that the right side of (\dagger) has all partial derivatives $\partial / \partial y^k$ equal to 0. But

$$\begin{aligned} \frac{\partial}{\partial y^k} \left(\sum_{i,j} g^{ij} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} \right) &= \sum_{i,j} \frac{\partial g^{ij}}{\partial y^k} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} \\ &\quad + \sum_{i,j} g^{ij} \frac{\partial^2 x^\mu}{\partial y^i \partial y^k} \frac{\partial x^\nu}{\partial y^j} + \sum_{i,j} g^{ij} \frac{\partial x^\mu}{\partial y^i} \frac{\partial^2 x^\nu}{\partial y^j \partial y^k} \\ &= \sum_{i,j} \frac{\partial g^{ij}}{\partial y^k} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} + \sum_{i,j} g^{ij} \sum_{\gamma} \Gamma_{ik}^{\gamma} \frac{\partial x^\mu}{\partial y^\gamma} \frac{\partial x^\nu}{\partial y^j} \\ &\quad + \sum_{i,j} g^{ij} \sum_{\gamma} \Gamma_{jk}^{\gamma} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^\gamma} \quad \text{by } (\dagger). \end{aligned}$$

Switching some indices, we thus have

$$\begin{aligned} \frac{\partial}{\partial y^k} \left(\sum_{i,j} g^{ij} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} \right) &= \sum_{i,j} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} \left(\frac{\partial g^{ij}}{\partial y^k} + \sum_{\gamma} g^{\gamma j} \Gamma_{\gamma k}^i + \sum_{\gamma} g^{i \gamma} \Gamma_{\gamma k}^j \right) \\ &= 0 \quad \text{by } (**). \end{aligned}$$

This completes the proof of the theorem. ♦

As a brief review of the proof, we note that

Step 1 uses the *integrability conditions*, $R = 0$, to obtain certain forms $\sum_i h^{(\alpha)}_i dy^i$, with any desired initial conditions;

Step 2 uses *symmetry of the Christoffel symbols*, $\Gamma_{ij}^k = \Gamma_{ji}^k$, to prove that $\sum_i h^{(\alpha)}_i dy^i = dx^\alpha$ for some x^α ;

Step 3 uses the *definition of the Christoffel symbols* $[ij, k]$ to prove that the vectors $\partial/\partial x^\alpha$ are orthonormal.

Despite its length, the proof is essentially a straightforward application of the integrability conditions for partial differential equations. As Riemann says, at the end of the section in Part C, “Given an acquaintance with the traditional methods, it is demonstrated without difficulty that these . . . conditions, when they are satisfied, suffice.”

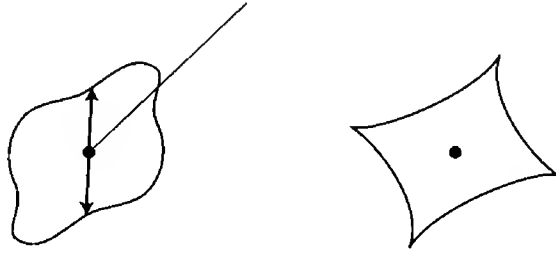
We have thus proved one special case of Riemann’s assertion that the curvature determines the metric. We will not return to the more general assertion until Chapter 7, for our immediate task will be to begin systematizing all the results which have been uncovered so far.

ADDENDUM FINSLER METRICS

A **Minkowski metric** on a vector space V is a function $F: V \rightarrow \mathbb{R}$ such that

$$\begin{aligned} F(v) &> 0 && \text{for all } v \neq 0 \\ F(\lambda v) &= |\lambda|F(v). \end{aligned}$$

Clearly F is completely determined by its “unit sphere” $\{v : F(v) = 1\}$; the unit sphere is symmetric with respect to $0 \in V$, and intersects every ray through 0 exactly once. Moreover, any such set is clearly the unit sphere for some F . The



function F is never C^∞ at 0 , but F^2 may be, in which case we will simply say that F is C^∞ . The general metric which Riemann mentions is essentially an assignment of a C^∞ Minkowski metric F_p to each tangent space M_p , in such a way that F_p varies smoothly with p . We will call such an assignment simply a “metric” on M .

If $c: [a, b] \rightarrow M$ is a curve in a manifold M with such a metric, then we can define the **length** of c to be

$$\int_a^b F_{c(t)}(c'(t)) dt.$$

If $p: [\bar{a}, \bar{b}] \rightarrow [a, b]$ is an increasing diffeomorphism, and we denote $F_{c(t)}(c'(t))$ by $g(t)$, then

$$\begin{aligned} \text{length of } c \circ p &= \int_{\bar{a}}^{\bar{b}} F_{c(p(t))}((c \circ p)'(t)) dt \\ &= \int_{\bar{a}}^{\bar{b}} F_{c(p(t))}(p'(t) \cdot c'(p(t))) dt \\ &= \int_{p^{-1}(a)}^{p^{-1}(b)} p'(t) g(p(t)) dt \\ &= \int_a^b g(t) dt = \text{length of } c. \end{aligned}$$

The same result clearly holds if p is decreasing; thus, the length of a curve is independent of parameterization. It is to insure this result that we require our metric to satisfy $F(\lambda v) = |\lambda| \cdot F(v)$.

Although a C^∞ metric F on a manifold M is not a tensor, it can be used to construct a tensor on the manifold TM . To do this, we first consider a C^∞ function $f: V \rightarrow \mathbb{R}$ on a vector space V . For any two vectors $v, w \in V$ we can form the second derivative

$$f_{**}(v)(w) = \left. \frac{d^2}{dt^2} \right|_{t=0} f(v + tw);$$

this is a sort of second order directional derivative at v . If v_1, \dots, v_n is a basis for V , and $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$\phi(a^1, \dots, a^n) = f\left(\sum_{i=1}^n a^i v_i\right),$$

then

$$f_{**}\left(\sum_{i=1}^n b^i v_i\right)\left(\sum_{i=1}^n c^i v_i\right) = \sum_{i,j} \frac{\partial^2 \phi}{\partial x^i \partial x^j}(b) \cdot c^i c^j.$$

The map $f_{**}(v): V \rightarrow \mathbb{R}$ is called the **Hessian** of f at $v \in V$. When F is a Minkowski metric, it is clear from the definition that

$$(F^2)_{**}(v)(v) = 2[F(v)]^2.$$

[This Hessian may be compared with the Hessian f_{**} defined in Problem I.5-17 for a function $f: M \rightarrow \mathbb{R}$, at a point $p \in M$ where $f_*p = 0$. The latter is a bilinear function on M_p , whereas the present $f_{**}(v)$ is a quadratic function on V . The associated bilinear function is easily seen to be

$$(w_1, w_2) \mapsto \left. \frac{\partial^2}{\partial s \partial t} \right|_{(s,t)=0} f(v + sw_1 + tw_2),$$

and in terms of a basis it is given by

$$\left(\sum_{i=1}^n c^i v_i, \sum_{i=1}^n d^i v_i\right) \mapsto \sum_{i,j} \frac{\partial^2 \phi}{\partial x^i \partial x^j} b \cdot c^i d^j,$$

the same formula which occurs in Problem I.5-17. Our Hessian is defined even at points where $f_* \neq 0$ because we are working with a vector space, and

identifying it with its tangent space at v ; this amounts to saying that we are considering only linear changes of coordinates, all of which leave the quantity defined by this formula invariant.]

Now if M is a manifold, with a C^∞ metric F , and $v \in M_p$, then the tangent space $(TM)_v$ of TM at v can be identified with M_p ; in accordance with this identification, we denote a vector in $(TM)_v$ by w_v . We can now define a tensor \mathcal{F} on TM by

$$\mathcal{F}(w_v) = \frac{1}{2}(F_p^2)_{**}(v)(w).$$

If F is the norm $\| \cdot \|$ associated with a Riemannian metric $\langle \cdot, \cdot \rangle$, then it is easy to see that

$$\mathcal{F}(w_v) = \langle v, w \rangle_p,$$

and in general we always have

$$[F(v_p)]^2 = \mathcal{F}(v_v).$$

If x is a coordinate system on M , and $(x \circ \pi, \dot{x})$ is the corresponding coordinate system on TM (defined on pg. I.81), then

$$\mathcal{F} = \sum_{i,j=1}^n g_{ij} d\dot{x}^i \cdot d\dot{x}^j,$$

where

$$g_{ij}(v_p) = \frac{1}{2} \frac{\partial^2 (F_p^2)}{\partial \dot{x}^i \partial \dot{x}^j}(v_p).$$

Classically, one dealt only with the functions g_{ij} , defined by this formula, and checked that the function

$$\sum_{i=1}^n b^i \frac{\partial}{\partial x^i} \Big|_p \mapsto \sum_{i,j=1}^n g_{ij} b^i b^j$$

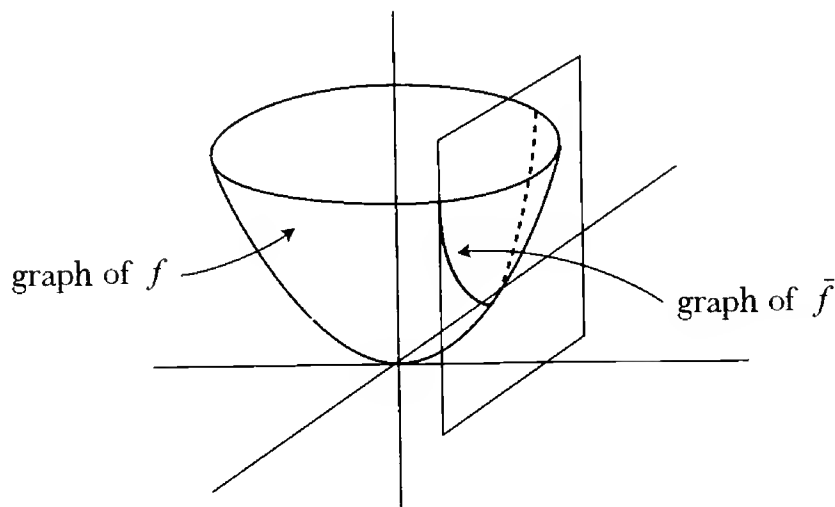
was independent of the coordinate system x .

A **Finsler metric** is defined to be a metric F such that the quadratic function $\mathcal{F}: (TM)_v \rightarrow \mathbb{R}$ is positive definite for all $v \in TM$. In terms of a coordinate system x , this means that

$$(g_{ij}(v_p)) = \frac{1}{2} \left(\frac{\partial^2 (F_p^2)}{\partial \dot{x}^i \partial \dot{x}^j}(v_p) \right) \quad \text{is a positive definite matrix.}$$

To interpret this condition geometrically, we consider once again a function $f: V \rightarrow \mathbb{R}$. Suppose that for all v we have $f_{**}(v)(w) > 0$ for all $w \neq 0$.

Then for all v , and each straight line through v , the function \bar{f} obtained by restricting f to this line has an everywhere positive second derivative. By a



standard theorem (*Calculus*, pg. 220) this means that the function \bar{f} is convex (the set of points above or on its graph is a convex subset of the plane). It is easy to see that consequently the function f itself must be convex (the set of points on or above its graph is a convex subset of $V \times \mathbb{R}$). Conversely, convexity of f implies that $f_{**}(v)(w) \geq 0$. Our stronger condition $f_{**}(v)(w) > 0$ might be called “very-strict-convexity of f ”.

When we apply this to a Finsler metric, we see that each function F_p^2 is convex on M_p , so that each $C_p = \{(v, r) \in M_p \times \mathbb{R} : r \geq F_p^2(v)\}$ is a convex subset of $M_p \times \mathbb{R}$. It is easy to see that this is equivalent to convexity of the “unit ball” $\{w : F_p(w) \leq 1\}$, which may be regarded as the intersection of C_p and the hyperplane $r = 1$ in $M_p \times \mathbb{R}$. Note that convexity of the unit ball is equivalent to the “triangle inequality”

$$F_p(w_1 + w_2) \leq F_p(w_1) + F_p(w_2).$$

In general, a function $\| \cdot \| : V \rightarrow \mathbb{R}$ on a finite dimensional space V is called a **Banach space norm** if

$$\begin{aligned} \|v\| &> 0 && \text{for } v \neq 0 \\ \|\lambda v\| &= |\lambda| \cdot \|v\| \\ \|v + w\| &\leq \|v\| + \|w\| \end{aligned}$$

(if V is infinite dimensional, the definition is more involved). So a Finsler metric on M is a C^∞ Banach space norm on each M_p , varying smoothly with p (and with the unit balls on each M_p satisfying a very-strict-convexity requirement).

Although we will not develop the theory of Finsler, or more general, metrics here, we will mention a few facts. In the case of a Finsler metric, since the matrix $(g_{ij}(v_p))$ is non-degenerate, we can define $g^{ij}(v_p)$ so that

$$\sum_{j=1}^n g^{ij}(v_p) \cdot g_{jk}(v_p) = \delta_k^i.$$

The reader may seek an invariant description of the $g^{ij}(v_p)$. We can also define the symbols

$$[ij, k] = \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right)$$

$$\Gamma_{ij}^k = \sum_{l=1}^n g^{kl} [ij, l],$$

as before, except that they are now functions on TM . It turns out that the critical paths for length are curves $c: [a, b] \rightarrow M$ which, when parameterized by arclength, satisfy

$$\frac{d^2 x^k(c(s))}{ds^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(c'(s)) \cdot \frac{d\dot{x}^i(c(s))}{ds} \frac{d\dot{x}^j(c(s))}{ds} = 0.$$

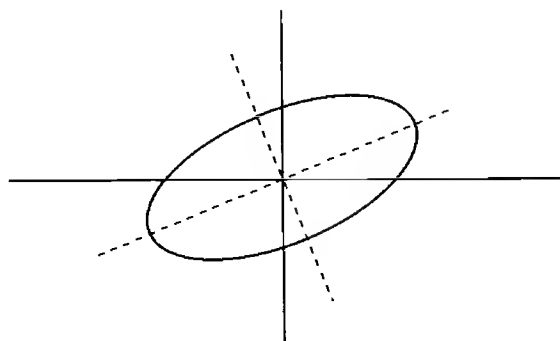
It also turns out, just as in the Riemannian case, that sufficiently small pieces of these critical paths are actually paths of shortest length. However, this is false for more general Minkowski metrics.

We shall not pursue the subject of Finsler metrics much further, but we will add some remarks about Minkowski metrics F on a vector space V . A Minkowski metric F can be used to define a “distance” function on $V \times V$, by $(v, w) \mapsto F(w - v)$ (however, it is easy to see that this distance function satisfies the triangle inequality, and is consequently a metric, if and only if F is a Banach space norm on V). This is just the procedure by which, in analytic geometry, we define the distance between two points in \mathbb{R}^n ; in this case, we choose $F(x) = (\sum (x^i)^2)^{1/2}$, motivated, of course, by the Pythagorean Theorem. After the Pythagorean Theorem has been incorporated into our definition of distance in this way, it is interesting to ask what content, if any, remains to this theorem. The answer is, that the Pythagorean Theorem has been declared true only for right triangles with sides parallel to the axes, but remains true for all right triangles. This is because, when $F(x) = (\sum (x^i)^2)^{1/2}$, the isometries of \mathbb{R}^n are transitive on the unit sphere. That is to say, if p and q are in the unit

sphere, then there is a linear transformation $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ with $\phi(p) = q$ and $F(\phi(x)) = F(x)$ for all x .

This same transitivity property holds for any F which is the norm $\| \cdot \|$ associated to a positive definite inner product $\langle \cdot, \cdot \rangle$ on an n -dimensional vector space V , since there is an isomorphism $f: \mathbb{R}^n \rightarrow V$ such that $f^*\langle \cdot, \cdot \rangle$ is the usual inner product on \mathbb{R}^n (Theorem I.9-3). It turns out that this property actually characterizes the Minkowski metrics F which arise from inner products. To prove this, we need an auxiliary concept, and a result from linear algebra.

An **ellipsoid** on a vector space V is a set of the form $\{v \in V : \langle v, v \rangle \leq 1\}$ for some positive definite inner product $\langle \cdot, \cdot \rangle$. In particular, consider such an inner product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^n , which also has its standard inner product $\langle \cdot, \cdot \rangle$. The ellipsoid $\{v \in V : \langle v, v \rangle \leq 1\}$ really looks like an ellipsoid, because of the



following well-known result:

14. PROPOSITION (EXISTENCE OF PRINCIPAL AXES). If $\langle \cdot, \cdot \rangle$ is any positive definite inner product on \mathbb{R}^n , then there is a basis for \mathbb{R}^n which is orthonormal for $\langle \cdot, \cdot \rangle$ and also *orthogonal* with respect to $\langle \cdot, \cdot \rangle$.

PROOF. For each $x \in \mathbb{R}^n$, the map $y \mapsto \langle x, y \rangle$ is a linear functional, so there is a unique $Tx \in \mathbb{R}^n$ such that

$$\langle Tx, y \rangle = \langle x, y \rangle \quad \text{for all } y.$$

It is easy to see that $x \mapsto Tx$ is a linear transformation $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$. Moreover,

$$\langle Tx, y \rangle = \langle x, y \rangle = \langle y, x \rangle = \langle Ty, x \rangle = \langle x, Ty \rangle,$$

so T is self-adjoint with respect to $\langle \cdot, \cdot \rangle$. Thus T has a basis x_1, \dots, x_n of eigenvalues, $Tx_i = \lambda_i x_i$, and the x_i can be picked orthogonal with respect

to $\langle \cdot, \cdot \rangle$. Now

$$\begin{aligned}\lambda_j \langle x_i, x_j \rangle &= \lambda_j \langle T x_i, x_j \rangle = \lambda_i \lambda_j \langle x_i, x_j \rangle \\ &= \lambda_i \langle x_i, T x_j \rangle \\ &= \lambda_i \langle x_i, x_j \rangle.\end{aligned}$$

So $\langle x_i, x_j \rangle = 0$ if $\lambda_i \neq \lambda_j$. On the other hand, if two or more x_i have the same λ_i , then in the m -dimensional subspace which they span we can pick m eigenvectors which are orthogonal with respect to $\langle \cdot, \cdot \rangle$. So we can assume that $\langle x_i, x_j \rangle = 0$ for $i \neq j$. Now we just normalize each x_i with respect to $\langle \cdot, \cdot \rangle$. ♦

We now use this result to prove the basic lemma for our main assertion.

15. LEMMA. Let B be a bounded neighborhood of 0 in an n -dimensional vector space V . Then among all ellipsoids containing B there exists a unique one of smallest volume. (We assign a volume to the ellipsoids by choosing an isomorphism of V with \mathbb{R}^n . Choosing a different isomorphism clearly does not change the property of having the “smallest volume”.)

PROOF. First we prove existence. We might as well assume that $V = \mathbb{R}^n$. Moreover, we can assume that B is closed, since an ellipsoid containing B also contains \bar{B} . Choose r and R so that

$$\{x \in \mathbb{R}^n : |x| \leq r\} \subset B \subset \{x \in \mathbb{R}^n : |x| \leq R\}.$$

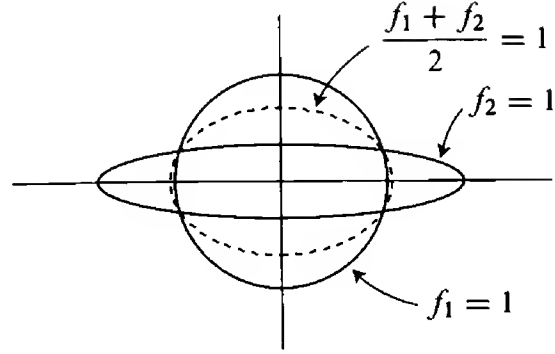
Every orthogonal basis $b = (v_1, \dots, v_n)$ of \mathbb{R}^n determines an ellipsoid $E(b)$ with principal axes v_1, \dots, v_n [equivalently, b determines an inner product on \mathbb{R}^n , namely the one which makes v_1, \dots, v_n *orthonormal*]. To prove existence it obviously suffices to consider only ellipsoids of volume $\leq \alpha R^n$, where α is the volume of the unit ball. Now the principal axes v_1, \dots, v_n of any ellipsoid containing B must have lengths $a_i \geq r$. Consequently, if this ellipsoid has volume $\leq \alpha R^n$, so that $\prod_i a_i \leq R^n$, then each $a_i \leq R^n / r^{n-1}$.

Consider the set $\{b = (v_1, \dots, v_n) : B \subset E(b) \text{ and length } v_i \leq R^n / r^{n-1}\}$. This is a compact subset of the n -fold product $\mathbb{R}^n \times \dots \times \mathbb{R}^n$. Hence $b \mapsto \text{volume } E$ takes on its minimum on this set. This proves existence.

Now consider two different ellipsoids containing B , with the same volume. Choose an isomorphism of V with \mathbb{R}^n which makes the first of these ellipsoids correspond to the ordinary unit ball $\{x \in \mathbb{R}^n : f_1(x) \leq 1\}$, where $f_1(x) = \sum (x^i)^2$. Proposition 14 shows that after a rotation of the axes, the second of

the ellipsoids corresponds to $\{x \in \mathbb{R}^n : f_2(x) \leq 1\}$, where

$$f_2(x) = a_1(x^1)^2 + \cdots + a_n(x^n)^2.$$



The volume of this ellipsoid is $\prod_{i=1}^n 1/\sqrt{a_i}$ times the volume of the unit ball. Since the two ellipsoids are assumed to have the same volume, this means that

$$\prod_{i=1}^n a_i = 1.$$

Now consider the ellipsoid

$$E = \left\{ x \in \mathbb{R}^n : \frac{f_1 + f_2}{2}(x) \leq 1 \right\}.$$

Clearly E also contains B . Now the semi-axes of E have length

$$\frac{1}{\sqrt{\frac{1+a_i}{2}}},$$

so the volume of E is the volume of the unit ball times

$$\prod_{i=1}^n \frac{1}{\sqrt{\frac{1+a_i}{2}}}.$$

Recall that for $a, b > 0$ we have

$$\sqrt{ab} \leq \frac{a+b}{2} \quad \text{with equality if and only if } a = b.$$

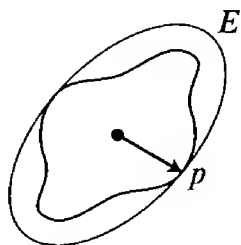
Consequently,

$$\prod_{i=1}^n \frac{1}{\sqrt{\frac{1+a_i}{2}}} \leq \prod_{i=1}^n \frac{1}{\sqrt[4]{a_i}} = \frac{1}{\sqrt[4]{\prod_{i=1}^n a_i}} = 1,$$

and strict equality holds if some $a_i \neq 1$, i.e., if the original two ellipsoids are different. This means that if two different ellipsoids containing B have the same volume, then there is another ellipsoid containing B with smaller volume. This clearly proves uniqueness of the ellipsoid with smallest volume. ♦

16. THEOREM. Let $F: V \rightarrow \mathbb{R}$ be a continuous Minkowski metric on an n -dimensional vector space V . Suppose that for all p and q in the unit sphere $\{v \in V : F(v) = 1\}$, there is a linear transformation $\phi: V \rightarrow V$ such that $\phi(p) = q$ and $F(\phi(v)) = F(v)$ for all $v \in V$. Then F is the norm determined by some positive definite inner product.

PROOF. Let $B = \{v : F(v) \leq 1\}$, and let E be the unique ellipsoid containing B of smallest volume. Clearly, there must be some point p with $F(p) = 1$ and $p \in \text{boundary } E$. Let q be any other point with $F(q) = 1$, and $\phi: V \rightarrow V$



a linear transformation with $\phi(p) = q$ such that $F(\phi(v)) = F(v)$ for all $v \in V$. It follows easily from the latter property that $\phi(E) \supset B$. Moreover, $\phi(B) = B$, so ϕ is volume preserving. By uniqueness of the ellipsoid E , it follows that $\phi(E) = E$. Consequently, $q = \phi(p) \in \text{boundary } E$. In other words, every point q with $F(q) = 1$ is in boundary E . This means that $E = B$. ♦

CHAPTER 5

THE ABSOLUTE DIFFERENTIAL CALCULUS (THE RICCI CALCULUS); OR, THE DEBAUCH OF INDICES

Although Riemann essentially defined the curvature tensor, the (classical) notion of a tensor did not even exist in his time. The development of the “Calculus of Tensors” is due mainly to Ricci, and was carried out in the years 1887–1896; in 1901 he and his student Levi-Civita gave a detailed description in a memoir *Methods de calcul differential absolu et leurs applications*. In addition to a comprehensive use of tensors, what distinguished the Absolute Differential Calculus, and gave it its name, was an important construction which greatly simplified all the concepts of Riemannian geometry, especially the curvature tensor. Instead of checking that the horrible formula used to define R^i_{jkl} transforms correctly, we are going to check that an equally mysterious—but not quite so horrible—formula transforms correctly, and thus defines a tensor; then we will define the curvature tensor in terms of this one.

In 1869, in one of the earliest papers which took up Riemann’s ideas, Christoffel had already made the observation which Ricci made the basis for his calculus. Suppose that Y is a vector field, and

$$\begin{aligned} Y &= \sum_{j=1}^n \lambda^j \frac{\partial}{\partial x^j} \\ &= \sum_{j=1}^n \lambda'^j \frac{\partial}{\partial x'^j}. \end{aligned}$$

Consider the following symbols, defined in terms of the Christoffel symbols of the second kind for the coordinate systems x and x' :

$$\begin{aligned} \lambda^j_{;h} &= \frac{\partial \lambda^j}{\partial x^h} + \sum_{v=1}^n \lambda^v \Gamma^j_{hv} \\ \lambda'^j_{;h} &= \frac{\partial \lambda'^j}{\partial x'^h} + \sum_{v=1}^n \lambda'^v \Gamma'^j_{hv}. \end{aligned}$$

It is easy to check that for these symbols—the sums of partial derivatives of the λ^j and certain linear combinations of them—we have

$$\lambda'^{\alpha}_{;\beta} = \sum_{h,j=1}^n \lambda^j_{;h} \frac{\partial x^h}{\partial x'^{\beta}} \frac{\partial x'^{\alpha}}{\partial x^j}.$$

For the calculation we use the formula on page 189 together with the formula

$$\lambda'^{\alpha} = \sum_{j=1}^n \lambda^j \frac{\partial x'^{\alpha}}{\partial x^j};$$

when we compute $\partial \lambda'^{\alpha} / \partial x'^{\beta}$, the extra terms involving second partial derivatives of the x'^{α} just cancel out with the extra terms in the transformation formula for the Christoffel symbols. This calculation shows that there is a certain tensor field of type $\binom{1}{1}$ which equals

$$\sum_{h,j=1}^n \lambda^j_{;h} dx^h \otimes \frac{\partial}{\partial x^j}$$

in the x coordinate system. In classical terms, “if the λ^j transform like a tensor of type $\binom{0}{1}$, then the $\lambda^j_{;h}$ transform like a tensor of type $\binom{1}{1}$ ”. Similarly, if the λ_i transform like a tensor of type $\binom{1}{0}$, then the quantities

$$\lambda_{i;h} = \frac{\partial \lambda_i}{\partial x^h} - \sum_{v=1}^n \lambda_v \Gamma_{hi}^v$$

are easily seen to transform like a tensor of type $\binom{2}{0}$, so that there is a tensor of type $\binom{2}{0}$ which equals

$$\sum_{i,h=1}^n \lambda_{i;h} dx^i \otimes dx^h$$

in the x coordinate system. Generally, given a tensor of type $\binom{k}{l}$, with components

$$A_{i_1 \dots i_k}^{j_1 \dots j_l},$$

there is a new tensor, of type $\binom{k+1}{l}$, with components

$$\begin{aligned} A_{i_1 \dots i_k;h}^{j_1 \dots j_l} &= \frac{\partial A_{i_1 \dots i_k}^{j_1 \dots j_l}}{\partial x^h} + \sum_{s=1}^l \sum_{v=1}^n A_{i_1 \dots i_k}^{j_1 \dots j_{s-1} v j_{s+1} \dots j_l} \Gamma_{hv}^{j_s} \\ &\quad - \sum_{r=1}^k \sum_{v=1}^n A_{i_1 \dots i_{r-1} v i_{r+1} \dots i_k}^{j_1 \dots j_l} \Gamma_{hv}^{i_r}. \end{aligned}$$

The proof is again just an enormous calculation. At the other extreme from this most general case is the special case of a tensor field of type $\binom{0}{0}$ on M , i.e., a function $f: M \rightarrow \mathbb{R}$. Here we simply define

$$f_{;h} = \frac{\partial f}{\partial x^h},$$

so that in this case at least the identity of the tensor field

$$\sum_{h=1}^n f_{;h} dx^h = \sum_{h=1}^n \frac{\partial f}{\partial x^h} dx^h$$

is no mystery—it is just df .

The tensor of type $\binom{k+1}{l}$ which we thus obtain from a tensor A of type $\binom{k}{l}$ is called the **covariant derivative** of A , since it is covariant of one order greater than A . It is also called the “absolute derivative” of A ; here the word “absolute” means that it doesn’t depend on a particular coordinate system. Various notations for the covariant derivative are encountered—one sometimes sees $\lambda_{i|h}^j$, or $\lambda_{i,h}^j$, or even λ_{ih}^j . Use of a semi-colon should avoid all possible confusion!

A partial answer to the question “What does the covariant derivative really mean?” is given by the following observation.

1. PROPOSITION. Let x be a Riemannian normal coordinate system at the point $p \in M$, and A a tensor of type $\binom{k}{l}$ on M , with

$$A = \sum A_{i_1 \dots i_k}^{j_1 \dots j_l} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}}.$$

Then the components *at the point* p of the covariant derivative of A are just the ordinary partial derivatives

$$(*) \quad A_{i_1 \dots i_k; h}^{j_1 \dots j_l}(p) = \frac{\partial A_{i_1 \dots i_k}^{j_1 \dots j_l}}{\partial x^h}(p).$$

PROOF. In a Riemannian normal coordinate system at p , the Christoffel symbols $[ij, k]$, and hence also the Γ_{ij}^k , are all 0 at p . ♦

Proposition 1 can even be used to *define* covariant derivatives. The Riemannian normal coordinate systems at p are a natural set of coordinate systems, determined by the metric $\langle \ , \ \rangle$; any two such systems at p differ only by an

element of $O(n)$, and it is easy to see from this that a definition of the covariant derivative by means of $(*)$ would not depend on which one was picked. With a little work, one could then deduce the general expression for the components of the covariant derivative of A in any coordinate system. In succeeding chapters we will be giving still other interpretations of the covariant derivative.

The operation of covariant differentiation obeys many rules analogous to those for ordinary differentiation.

2. PROPOSITION. Covariant differentiation is a derivation and commutes with sums and contractions. For example, if A and B are tensors of type $\binom{2}{2}$ and C is a tensor of type $\binom{1}{1}$, then

$$\begin{aligned}(A_{i_1 i_2}^{j_1 j_2} C_{i_3}^{j_3})_{;h} &= A_{i_1 i_2;h}^{j_1 j_2} C_{i_3}^{j_3} + A_{i_1 i_2}^{j_1 j_2} C_{i_3;h}^{j_3} \\ (A_{i_1 i_2}^{j_1 j_2} + B_{i_1 i_2}^{j_1 j_2})_{;h} &= A_{i_1 i_2;h}^{j_1 j_2} + B_{i_1 i_2;h}^{j_1 j_2} \\ \left(\sum_{v=1}^n A_{i_1 v}^{v j_2} \right)_{;h} &= \sum_{v=1}^n A_{i_1 v;h}^{v j_2}.\end{aligned}$$

Consequently, we have, for example,

$$\left(\sum_{v=1}^n A_{v i_2}^{j_1 j_2} C_{i_3}^v \right)_{;h} = \sum_{v=1}^n A_{v i_2;h}^{j_1 j_2} C_{i_3}^v + A_{v i_2}^{j_1 j_2} C_{i_3;h}^v.$$

PROOF. Compute. Because of Proposition 1, the computations become trivial if one uses Riemannian normal coordinates. ♦

There is one tensor which we have on any manifold, the identity map of each M_p into itself, with coordinates δ_i^j in any coordinate system. For the covariant derivative we have

$$\begin{aligned}\delta_{i;h}^j &= \frac{\partial \delta_i^j}{\partial x^h} + \sum_{v=1}^n \delta_i^v \Gamma_{hv}^j - \sum_{v=1}^n \delta_v^j \Gamma_{hi}^v \\ &= \Gamma_{hi}^j - \Gamma_{hi}^j \\ &= 0,\end{aligned}$$

which shouldn't be very surprising (what else could it be?). Aside from these general formulas, there are two of crucial importance. The first of them is about the covariant derivative of the tensors $\langle \ , \ \rangle$ and $\langle \ , \ \rangle^*$.

3. PROPOSITION (RICCI'S LEMMA). The g_{ij} and g^{ij} behave like constants in covariant differentiation; that is,

$$\begin{aligned} g_{ij;k} &= 0 \\ g^{ij};_k &= 0. \end{aligned}$$

PROOF. The proof is, of course, a calculation. For the g_{ij} we have

$$\begin{aligned} g_{ij;k} &= \frac{\partial g_{ij}}{\partial x^k} - \sum_{v=1}^n g_{vj} \Gamma_{ki}^v - \sum_{v=1}^n g_{iv} \Gamma_{kj}^v \\ &= \frac{\partial g_{ij}}{\partial x^k} - [ik, j] - [jk, i] \\ &= 0, \end{aligned}$$

by equation (*) on page 196. Similarly, the second equation is equivalent to the following equation (**). It can also be obtained from the first equation and Proposition 2, since $\sum_j g^{ij} g_{jl} = \delta_l^i$ and $\delta_{l,k}^i = 0$. ♦

The second crucial formula involves “second order” covariant derivatives. If A is a tensor of type $\binom{k}{l}$, with components $A_{i_1 \dots i_k}^{j_1 \dots j_l}$, then the operation of covariant differentiation can be applied to the tensor B with components

$$B_{i_1 \dots i_k h}^{j_1 \dots j_l} = A_{i_1 \dots i_k; h}^{j_1 \dots j_l};$$

there results a tensor C with components

$$C_{i_1 \dots i_k h \eta}^{j_1 \dots j_l} = B_{i_1 \dots i_k h; \eta}^{j_1 \dots j_l} = (A_{i_1 \dots i_k; h}^{j_1 \dots j_l})_{; \eta}.$$

These components are denoted by

$$A_{i_1 \dots i_k; h; \eta}^{j_1 \dots j_l} \quad \text{or simply} \quad A_{i_1 \dots i_k; h \eta}^{j_1 \dots j_l}.$$

For example, if we start with a function f [a tensor of type $\binom{0}{0}$], then

$$f_{;i} = \frac{\partial f}{\partial x^i}$$

and

$$\begin{aligned} f_{;ij} &= (f_{;i})_{;j} = \frac{\partial \left(\frac{\partial f}{\partial x^i} \right)}{\partial x^j} - \sum_{v=1}^n \frac{\partial f}{\partial x^v} \Gamma_{ji}^v \\ &= \frac{\partial^2 f}{\partial x^i \partial x^j} - \sum_{v=1}^n \frac{\partial f}{\partial x^v} \Gamma_{ji}^v. \end{aligned}$$

Notice that

$$f_{;ij} = f_{;ji}$$

by symmetry of the Γ_{ij}^ν . The same result definitely does not hold for other tensors; instead we have the following basic result.

4. PROPOSITION (RICCI'S IDENTITIES). If the λ^i and λ_i are components of tensors of type $\binom{0}{1}$ and $\binom{1}{0}$, respectively, then

$$\lambda^i_{;jk} - \lambda^i_{;kj} = - \sum_{l=1}^n \lambda^l R^i_{ljk}$$

$$\lambda_{i;jk} - \lambda_{i;kj} = \sum_{l=1}^n \lambda_l R^l_{ijk},$$

where

$$R^i_{jkl} = \frac{\partial \Gamma^i_{lj}}{\partial x^k} - \frac{\partial \Gamma^i_{kj}}{\partial x^l} + \sum_{\mu=1}^n \Gamma^{\mu}_{lj} \Gamma^i_{k\mu} - \Gamma^{\mu}_{kj} \Gamma^i_{l\mu}.$$

(There are similar identities for tensors of type $\binom{k}{l}$, but we will ignore them.*)

PROOF. Compute. ♦

[The second identity is a consequence of the first, for if we are given λ_i and define

$$\lambda^i = \sum_{\alpha=1}^n g^{i\alpha} \lambda_{\alpha}, \quad \text{so that} \quad \lambda_i = \sum_{\alpha=1}^n g_{i\alpha} \lambda^{\alpha},$$

then

$$\begin{aligned} \lambda_{i;jk} - \lambda_{i;kj} &= \left(\sum_{\alpha=1}^n g_{i\alpha} \lambda^{\alpha} \right)_{;jk} - \left(\sum_{\alpha=1}^n g_{i\alpha} \lambda^{\alpha} \right)_{;kj} \\ &= \sum_{\alpha=1}^n g_{i\alpha} (\lambda^{\alpha}_{;jk} - \lambda^{\alpha}_{;kj}) \quad \text{using Proposition 2} \end{aligned}$$

* For those who cannot bear to be left in ignorance, the general Ricci identity is

$$A^{j_1 \dots j_l}_{i_1 \dots i_k; h\eta} - A^{j_1 \dots j_l}_{i_1 \dots i_k; \eta h} = \sum_{r=1}^k \sum_{\nu=1}^n A^{j_1 \dots j_l}_{i_1 \dots i_{r-1} \nu i_{r+1} \dots i_k} R^{\nu}_{i_r h \eta} - \sum_{s=1}^l \sum_{\nu=1}^n A^{j_1 \dots j_{s-1} \nu j_{s+1} \dots j_l}_{i_1 \dots i_k} R^{j_s}_{\nu h \eta}.$$

$$\begin{aligned}
 &= - \sum_{\alpha=1}^n \sum_{l=1}^n g_{i\alpha} \lambda^l R^{\alpha}{}_{ljk} \\
 &= - \sum_{l=1}^n \lambda^l R_{iljk} && \text{(by definition of } R_{iljk}\text{)} \\
 &= \sum_{l=1}^n \lambda^l R_{lijk} && \text{by Proposition 4-10} \\
 &= \sum_{l=1}^n \sum_{\alpha=1}^n g_{l\alpha} \lambda^l R^{\alpha}{}_{ijk} \\
 &= \sum_{\alpha=1}^n \lambda_{\alpha} R^{\alpha}{}_{ijk},
 \end{aligned}$$

and similarly the first identity is a consequence of the second.]

5. COROLLARY. The $R^i{}_{jkl}$ are the components of a tensor of type $\binom{3}{1}$.

PROOF. Let Z be a vector field, with

$$Z = \sum_{i=1}^n \lambda^i \frac{\partial}{\partial x^i},$$

and let A be the tensor field with components $\lambda^i{}_{;jk}$. The first equation in Proposition 4 shows that

$$\begin{aligned}
 (*) \quad B\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) &= A\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) - A\left(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^j}\right) \\
 &= \sum_{i=1}^n (\lambda^i{}_{;jk} - \lambda^i{}_{;kj}) \frac{\partial}{\partial x^i} \\
 &= - \sum_{i=1}^n \left(\sum_{l=1}^n \lambda^l R^i{}_{ljk} \right) \frac{\partial}{\partial x^i} \\
 &= - \sum_{l=1}^n \lambda^l \left(\sum_{i=1}^n R^i{}_{ljk} \frac{\partial}{\partial x^i} \right).
 \end{aligned}$$

This shows, in particular, that $B(\partial/\partial x^j|_p, \partial/\partial x^k|_p)$ does not depend on the vector field Z , but on the vector $Z(p)$ alone. So if we define

$$R(X, Y)Z = -B(X, Y),$$

then R is a tensor of type $\binom{3}{1}$, and

$$\begin{aligned} B\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) &= -R\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) Z \\ &= -\sum_{l=1}^n \lambda^l R\left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\right) \frac{\partial}{\partial x^l}. \end{aligned}$$

Since this is true for all n -tuples $\{\lambda^l\}$, comparison with equation (*) shows that the components of R are indeed $R^i{}_{ljk}$. ♦

[Classically, this corollary would be deduced from the following general principle (compare Problem I.4-5(i)):

Suppose we are given a set of numbers $T^i{}_{ljk}$ for the coordinate system x , a set $T'^i{}_{ljk}$ for the coordinate system x' , etc. Suppose also that

$$C^i{}_{jk} = \sum_{l=1}^n \lambda^l T^i{}_{ljk}, \quad C'^i{}_{jk} = \sum_{l=1}^n \lambda'^l T'^i{}_{ljk}, \quad \text{etc.,}$$

are the components of a tensor for all tensors of type $\binom{0}{1}$ with components λ^l in the coordinate system x , and components λ'^l in the coordinate system x' , etc. Then $T^i{}_{ljk}$ are the components of a tensor.

The classical proof is by a calculation. We have

$$\begin{aligned} \sum_{\delta=1}^n \lambda'^{\delta} T'^{\alpha}{}_{\delta\beta\gamma} &= C'^{\alpha}{}_{\beta\gamma} = \sum_{i,j,k} C^i{}_{jk} \frac{\partial x'^{\alpha}}{\partial x^i} \frac{\partial x^j}{\partial x'^{\beta}} \frac{\partial x^k}{\partial x'^{\gamma}} \\ &= \sum_{i,j,k,l} \lambda^l T^i{}_{ljk} \frac{\partial x'^{\alpha}}{\partial x^i} \frac{\partial x^j}{\partial x'^{\beta}} \frac{\partial x^k}{\partial x'^{\gamma}} \\ &= \sum_{i,j,k,l,\delta} \lambda'^{\delta} T^i{}_{ljk} \frac{\partial x^l}{\partial x'^{\delta}} \frac{\partial x'^{\alpha}}{\partial x^i} \frac{\partial x^j}{\partial x'^{\beta}} \frac{\partial x^k}{\partial x'^{\gamma}}. \end{aligned}$$

Since this is true for arbitrary λ , we can choose all λ'^{δ} but one equal to zero; this gives the desired transformation formulas.]

Corollary 5 represents only one minor application of the Ricci identities. A more significant application is obtained when we consider a manifold with vanishing curvature tensor. In this case, we have

$$\begin{aligned} \lambda^i{}_{;jk} &= \lambda^i{}_{;kj} \\ \lambda_{i;jk} &= \lambda_{i;kj}. \end{aligned}$$

Thus, in manifolds with a vanishing curvature tensor (and only in such manifolds) the order of covariant differentiation is immaterial, so that in this respect covariant differentiation behaves like ordinary partial differentiation.* This enables us to give a more direct proof of

6. THEOREM (THE TEST CASE; SECOND VERSION). Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. As before, we assume we are in \mathbb{R}^n , and choose the standard coordinate system y^1, \dots, y^n around 0.

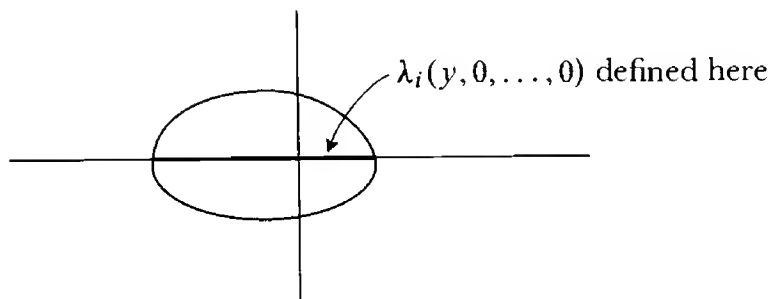
Step 1. We claim that we can find 1-forms $\eta = \sum \lambda_i dy^i$, with any desired initial value $\eta(0)$, satisfying

$$(*) \quad \lambda_{i;j} = 0 \quad i, j = 1, \dots, n.$$

To prove this, we begin by finding $\lambda_i(y, 0, \dots, 0)$, with the prescribed value for $y = 0$, and such that

$$(*)_1 \quad 0 = \lambda_{i;1}(y, 0, \dots, 0) = \frac{\partial \lambda_i(y, 0, \dots, 0)}{\partial y^1} - \sum_{v=1}^n \lambda_v(y, 0, \dots, 0) \Gamma_{1i}^v(y, 0, \dots, 0).$$

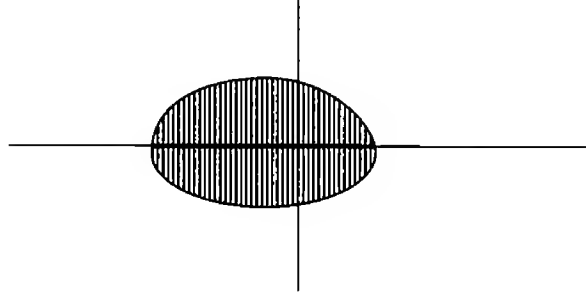
This just involves solving a set of ordinary differential equations, which is even linear, so that solutions exist for all y where the Γ 's are defined.



*At first sight, it might appear that this should always be the case, since covariant differentiation is the same as ordinary partial differentiation in a Riemannian normal coordinate system around p . However, the relation $\lambda^i_{;j} = \partial \lambda^i / \partial x^j$ holds only at p , so generally $\lambda^i_{;jk}(p) \neq \partial^2 \lambda^i / \partial x^j \partial x^k(p)$.

Next we find $\lambda_i(y_1, y, 0, \dots, 0)$, with the initial values $\lambda_i(y_1, 0, \dots, 0)$ just obtained, satisfying

$$\begin{aligned}
 (*_2) \quad 0 &= \lambda_{i;2}(y_1, y, 0, \dots, 0) \\
 &= \frac{\partial \lambda_i(y_1, y, 0, \dots, 0)}{\partial y^2} \\
 &\quad - \sum_{v=1}^n \lambda_v(y_1, y, 0, \dots, 0) \Gamma_{2i}^v(y_1, y, 0, \dots, 0).
 \end{aligned}$$



We continue in this way until we eventually obtain $\lambda_i(y_1, \dots, y_{n-1}, y)$ satisfying

$$(*_n) \quad 0 = \lambda_{i;n}(y_1, \dots, y_{n-1}, y).$$

We now claim that, in addition to the relation $\lambda_{i;1}(y, 0, \dots, 0) = 0$ given by $(*_1)$, we actually have

$$\lambda_{i;1}(y_1, y, 0, \dots, 0) = 0.$$

To see this, we first note that we have

$$\lambda_{i;21} = \lambda_{i;2;1} = \frac{\partial \lambda_{i;2}}{\partial y^1} - \sum_{v=1}^n \lambda_{v;2} \Gamma_{1i}^v - \sum_{v=1}^n \lambda_{i;v} \Gamma_{12}^v.$$

Since $0 = \lambda_{i;2}(y_1, y, 0, \dots, 0)$ by $(*_2)$, we obtain

$$\lambda_{i;21} = - \sum_{v=1}^n \lambda_{i;v} \Gamma_{12}^v \quad \text{at } (y_1, y, 0, \dots, 0).$$

Since $R = 0$, the Ricci identities then imply that

$$\lambda_{i;12} = \lambda_{i;21} = - \sum_{v=1}^n \lambda_{i;v} \Gamma_{12}^v \quad \text{at } (y_1, y, 0, \dots, 0),$$

i.e., that

$$\frac{\partial \lambda_{i;1}}{\partial y^2} - \sum_{v=1}^n \lambda_{v;1} \Gamma_{2i}^v - \sum_{v=1}^n \lambda_{1;v} \Gamma_{21}^v = - \sum_{v=1}^n \lambda_{1;v} \Gamma_{12}^v \quad \text{at } (y_1, y, 0, \dots, 0),$$

so that

$$\frac{\partial \lambda_{i;1}}{\partial y^2} - \sum_{v=1}^n \lambda_{v;1} \Gamma_{2i}^v = 0 \quad \text{at } (y_1, y, 0, \dots, 0).$$

Since we have the initial conditions $\lambda_{i;1}(y_1, 0, \dots, 0) = 0$, it is clear that the solution of this equation is just the desired one,

$$\lambda_{i;1}(y_1, y, 0, \dots, 0) = 0.$$

Proceeding in the same way, we next obtain

$$\lambda_{i;1}(y_1, y_2, y, 0, \dots, 0) = \lambda_{i;2}(y_1, y_2, y, 0, \dots, 0) = \lambda_{i;3}(y_1, y_2, y, 0, \dots, 0) = 0,$$

and, eventually,

$$0 = \lambda_{i;1} = \lambda_{i;2} = \dots = \lambda_{i;n} \quad \text{at } (y_1, \dots, y_{n-1}, y).$$

This completes the proof of the claim.

For $\alpha = 1, \dots, n$ we now choose $\eta^{(\alpha)} = \sum_i \lambda^{(\alpha)}_i dy^i$, so that

$$X_\alpha = (\lambda^{(\alpha)}_1(0), \dots, \lambda^{(\alpha)}_n(0))_0$$

are orthonormal with respect to $\langle \cdot, \cdot \rangle_0$.

Step 2. We claim that if $\lambda_{i;j} = 0$, then $\sum_i \lambda_i dy^i = dx$ for some function x . This is because we have

$$0 = \lambda_{i;j} = \frac{\partial \lambda_i}{\partial y^j} - \sum_{v=1}^n \lambda_v \Gamma_{ji}^v,$$

which shows that

$$\frac{\partial \lambda_i}{\partial y^j} = \frac{\partial \lambda_j}{\partial y^i}.$$

Now choose functions x^α with $\lambda^{(\alpha)}_i = \partial x^\alpha / \partial y^i$. As before, the x^α are a coordinate system in a neighborhood of 0.

Step 3. We claim that x is the desired coordinate system, i.e., that

$$\delta_{\mu\nu} = \sum_{i,j=1}^n g^{ij} \frac{\partial x^\mu}{\partial y^i} \frac{\partial x^\nu}{\partial y^j} = \sum_{i,j=1}^n g^{ij} \lambda^{(\mu)}_i \lambda^{(\nu)}_j.$$

As before, we just have to prove that the right side of this equation has all partial derivatives $\partial/\partial y^k$ equal to 0. But

$$\begin{aligned} \frac{\partial}{\partial y^k} \left(\sum_{i,j} g^{ij} \lambda^{(\mu)}_i \lambda^{(\nu)}_j \right) &= \left(\sum_{i,j} g^{ij} \lambda^{(\mu)}_i \lambda^{(\nu)}_j \right)_{;k} \\ &= \sum_{i,j} g^{ij}_{;k} \lambda^{(\mu)}_i \lambda^{(\nu)}_j + g^{ij} \lambda^{(\mu)}_{i;k} \lambda^{(\nu)}_j \\ &\quad + g^{ij} \lambda^{(\mu)}_i \lambda^{(\nu)}_{j;k} \\ &\quad \text{by Proposition 2} \\ &= 0 \text{ by Ricci's Lemma and equations (*). } \blacklozenge \end{aligned}$$

Comparing this proof of the Test Case with the first proof, we see that

Step 1 uses the *conditions* $R = 0$ to obtain the forms $\sum_i \lambda^{(\alpha)}_i dy^i$ satisfying (*). Instead of appealing to Theorem I.6-1, we essentially reprove this theorem; the Ricci identities make the proof almost as easy as the proof of Theorem I.6-0, the only complication being that the “mixed covariant derivatives” $\lambda_{i;12}$ depend on all $\lambda_{i;j}$, not just on $\lambda_{i;1}$.

Step 2, precisely the same as before, uses *symmetry of the Christoffel symbols*, $\Gamma_{ij}^k = \Gamma_{ji}^k$.

Step 3 uses the *definition of the Christoffel symbols* $[ij, k]$ to prove that the vectors $\partial/\partial x^\alpha$ are orthonormal. The proof is simpler because some of the calculations have been absorbed into the proof of Proposition 2, while the calculations using (**) on page 196 have been incorporated into Ricci's Lemma.

The absolute differential calculus turned out to be so useful (for many other applications besides the one just given) that it was soon exploited in the way all successful mathematical theories are—it was generalized. Notice that the

possibility of defining covariant derivatives depends only on the equation

$$(*) \quad \Gamma'^{\gamma}_{\alpha\beta} = \sum_{i,j,k} \Gamma^k_{ij} \frac{\partial x^i}{\partial x'^{\alpha}} \frac{\partial x^j}{\partial x'^{\beta}} \frac{\partial x'^{\gamma}}{\partial x^k} + \sum_{\mu=1}^n \frac{\partial^2 x^{\mu}}{\partial x'^{\alpha} \partial x'^{\beta}} \frac{\partial x'^{\gamma}}{\partial x^{\mu}};$$

it does not depend on the particular way that the Γ^k_{ij} are defined in terms of the g_{ij} . This observation suggests that we focus our attention on the transformation law (*) itself. Quantities which transform in this way are classically called *connections*. More precisely:

A **(classical) connection** on a manifold M is an assignment of n^3 numbers to each coordinate system, such that equation (*) holds between the n^3 numbers Γ^k_{ij} assigned to the coordinate system x and the n^3 numbers Γ'^k_{ij} assigned to the coordinate system x' .

Although this definition is exceedingly unappealing, classically it was motivated in the following way. If f is a function, then the quantities $\partial f / \partial x^i$ are the components of a tensor, but if λ^i are the components of a tensor, then the quantities $\partial \lambda^i / \partial x^j$ are not. If we attempt to construct a tensor by adding linear combinations of the λ^i , thus obtaining

$$\frac{\partial \lambda^i}{\partial x^j} + \sum_{\mu=1}^n \lambda^{\mu} \Gamma^i_{j\mu},$$

we find that these quantities are the components of a tensor provided that the Γ^k_{ij} transform according to (*).

Once we are given a connection (whatever in the world this connection may mean), we can imitate most of the work already done in this chapter for the special case where the Γ^k_{ij} are the Christoffel symbols, and in addition we can generalize other considerations from Riemannian geometry. What follows is a brief outline of this program.

We note first that, in contrast to the special case where the Γ^k_{ij} are the Christoffel symbols, a general connection need not satisfy $\Gamma^k_{ij} = \Gamma^k_{ji}$. However, it is easy to see that the quantities

$$T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}$$

are the components of a *tensor* T , the **torsion tensor** of the connection. (In the next chapter, we will see the reason for the term “connection”, but no one seems

to have a good explanation for the term “torsion” in this case.) A connection Γ_{ij}^k is called **symmetric** if the torsion tensor is zero. In this case, $T_{ij}^k = 0$ for every coordinate system, so $\Gamma_{ij}^k = \Gamma_{ji}^k$ for every coordinate system. Conversely, if $\Gamma_{ij}^k = \Gamma_{ji}^k$ in a set of coordinate systems which cover M , then the connection is symmetric.

The following result gives at least a little geometric significance to symmetry of a connection.

7. PROPOSITION. The torsion tensor T of a connection satisfies $T(p) = 0$ if and only if there is a coordinate system around p with

$$\Gamma_{ij}^k(p) = 0 \quad \text{for all } i, j, k.$$

PROOF. If $\Gamma_{ij}^k(p) = 0$, then $T(p) = 0$, so $T'^k_{ij} = 0$ in any coordinate system x' , which means that $\Gamma'^\gamma_{\alpha\beta} = \Gamma'^\gamma_{\beta\alpha}$.

Conversely, suppose that $\Gamma_{ij}^k(p) = \Gamma_{ji}^k(p)$ for all i, j, k in a coordinate system x . Define x'^k by

$$x'^k(q) = [x^k(q) - x^k(p)] + \frac{1}{2} \sum_{i,j=1}^n \Gamma_{ij}^k(p) [x^i(q) - x^i(p)] \cdot [x^j(q) - x^j(p)].$$

Using $\Gamma_{ij}^k(p) = \Gamma_{ji}^k(p)$, we compute that

$$\begin{aligned} \frac{\partial x'^k}{\partial x^l} &= \delta_l^k + \sum_{i=1}^n \Gamma_{il}^k(p) [x^i - x^i(p)] \\ \frac{\partial x'^k}{\partial x^l}(p) &= \delta_l^k. \end{aligned}$$

This shows that x' is a coordinate system in a neighborhood of p . Moreover,

$$\frac{\partial^2 x'^k}{\partial x^k \partial x^l}(p) = \Gamma_{il}^k(p).$$

Substituting into (*), we see that $\Gamma'^\gamma_{\alpha\beta}(p) = 0$. ♦

Proposition 7 becomes more significant when we introduce the concept of geodesics. In our treatment of Riemannian metrics, we defined geodesics as

critical points for the energy function. For a general connection Γ_{ij}^k , we can simply define a **geodesic** as a path γ satisfying

$$\frac{d^2\gamma^k}{dt^2} + \sum_{i,j=1} \Gamma_{ij}^k \circ \gamma \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0;$$

a calculation shows this condition does not depend on the coordinate system. The basic theorems on differential equations show that geodesics through p are uniquely determined by their tangent vectors $\gamma'(0) \in M_p$; we can thus define $\exp: M_p \rightarrow M$ as before.* We can also introduce “Riemannian normal coordinates” at p ; we choose any basis X_1, \dots, X_n for M_p , define $\chi: M_p \rightarrow \mathbb{R}^n$ by $\chi(\sum_{i=1}^n a^i X_i) = (a^1, \dots, a^n)$, and let $x = \chi \circ \exp^{-1}$. As in Proposition 4-1, we note that the geodesic $\gamma^k(t) = \xi^k t$ satisfies

$$\sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \xi^i \xi^j = 0,$$

so that

$$\sum_{i,j=1}^n \Gamma_{ij}^k(p) \xi^i \xi^j = 0 \quad \text{for all } n\text{-tuples } (\xi^1, \dots, \xi^n).$$

This implies that

$$\Gamma_{ij}^k(p) + \Gamma_{ji}^k(p) = 0;$$

if we also have $T(p) = 0$, then we deduce that $\Gamma_{ij}^k(p) = 0$.

We define covariant derivatives of tensors and the curvature tensor R for a connection by the formulas on pages 210 and 214. Notice that Proposition 1 holds for a *symmetric* connection, while Proposition 2 holds for any connection. Naturally, Proposition 3 has no analogue for general connections. For non-symmetric connections, the comparison of mixed covariant derivatives becomes a little more complicated. Recall that for a function $f: M \rightarrow \mathbb{R}$ we have

$$f_{;ij} = \frac{\partial^2 f}{\partial x^i \partial x^j} - \sum_{v=1}^n \frac{\partial f}{\partial x^v} \Gamma_{ji}^v.$$

When the connection is not symmetric, we can express $f_{;ij} - f_{;ji}$ in terms of the torsion tensor, which is also involved in the new Ricci identities.

*Many global results about geodesics for a Riemannian metric do *not* hold for more general connections; see Chapter 8, Addendum 2.

8. PROPOSITION. For a connection Γ_{ij}^k , with curvature tensor R and torsion tensor T , we have

$$\begin{aligned} f_{;ij} - f_{;ji} &= \sum_{v=1}^n \frac{\partial f}{\partial x^v} T_{ij}^v \\ \lambda^i_{;jk} - \lambda^i_{;kj} &= - \sum_{l=1}^n \lambda^l R^i_{ljk} + \sum_{l=1}^n \lambda^i_{;l} T^l_{jk} \\ \lambda_{i;jk} - \lambda_{i;kj} &= \sum_{l=1}^n \lambda_l R^l_{ijk} + \sum_{l=1}^n \lambda_{i;l} T^l_{jk}. \end{aligned}$$

PROOF. Compute, compute. ♦

(For all these relations it is, of course, extremely important that, in the definitions, care is given to the order of the subscripts in the Γ 's.)

Finally, we wish to consider the properties of the curvature tensor which are given in Proposition 4-10. Two of these have no analogue for a general connection, since they involve a metric, but there is an additional relation, involving the covariant derivative, which holds for any connection.

9. PROPOSITION. The curvature tensor for any connection satisfies the following identities:

$$(1) \quad R^i_{jkl} = -R^i_{jlk}$$

(2) (Bianchi's first identity)

$$\begin{aligned} R^i_{jkl} + R^i_{klj} + R^i_{ljk} \\ = (T^i_{kl;j} + T^i_{lj;k} + T^i_{jk;l}) + \sum_{\mu=1}^n (T^{\mu}_{jk} T^i_{\mu l} + T^{\mu}_{kl} T^i_{\mu j} + T^{\mu}_{lj} T^i_{\mu k}) \end{aligned}$$

(3) (Bianchi's second identity)

$$\begin{aligned} (R^h_{ijk;l} + R^h_{ikl;j} + R^h_{ilj;k}) \\ + \sum_{\mu=1}^n (T^{\mu}_{jk} R^h_{i\mu l} + T^{\mu}_{kl} R^h_{i\mu j} + T^{\mu}_{lj} R^h_{i\mu k}) = 0. \end{aligned}$$

In particular, if the connection is symmetric, then we have the much simpler relations

$$(2') \quad R^i_{jkl} + R^i_{klj} + R^i_{ljk} = 0$$

$$(3') \quad R^h_{ijk;l} + R^h_{ikl;j} + R^h_{ilj;k} = 0.$$

Classically, (3') alone is known as "Bianchi's identity".

PROOF. Equation (1) follows immediately from the definition. In the case of a symmetric connection, equation (2') is also easy to verify (we have already done it in Proposition 4-10). It is even simpler to verify if we use Riemannian normal coordinates at p ; in this case the definition (page 214) gives

$$R^i_{jkl}(p) = \frac{\partial \Gamma^i_{lj}}{\partial x^k}(p) - \frac{\partial \Gamma^i_{kj}}{\partial x^l}(p),$$

which yields (2') at once. Using Proposition 1, for a symmetric connection, we also obtain

$$R^h_{ijk;l}(p) = \frac{\partial^2 \Gamma^h_{ki}}{\partial x^l \partial x^j}(p) - \frac{\partial^2 \Gamma^h_{ji}}{\partial x^l \partial x^k}(p),$$

which gives (3').

The proof of (2) and (3) in general is considerably more complicated. To begin with, notice that $T^i_{\mu l} = -T^i_{l\mu}$, so

$$\sum_{\mu} T^{\mu}_{jk} T^i_{\mu l} = \sum_{\mu} \Gamma^{\mu}_{jk} T^i_{\mu l} + \sum_{\mu} \Gamma^{\mu}_{kj} T^i_{l\mu}.$$

We also have

$$T^i_{jk;l} = \frac{\partial T^i_{jk}}{\partial x^l} + \sum_{\mu} \Gamma^i_{l\mu} T^{\mu}_{jk} - \sum_{\mu} \Gamma^{\mu}_{lj} T^i_{\mu k} - \sum_{\mu} \Gamma^{\mu}_{lk} T^i_{j\mu}.$$

From these equations we see that the right side of (2) equals

$$\left(\frac{\partial T^i_{jk}}{\partial x^l} + \sum_{\mu} T^{\mu}_{jk} \Gamma^i_{l\mu} \right) + \text{the two terms obtained by cyclically permuting } j, k, l.$$

Using the definition of T^i_{jk} , this is easily seen to equal the left side of (2).

To prove (3), we note first that (1) gives

$$\sum_{\mu} T^{\mu}_{jk} R^h_{i\mu l} = \sum_{\mu} \Gamma^{\mu}_{jk} R^h_{i\mu l} + \Gamma^{\mu}_{kj} R^h_{i l\mu}.$$

We also have

$$R^h_{ijk;l} = \frac{\partial R^h_{ijk}}{\partial x^l} + \sum_{\mu} R^{\mu}_{ijk} \Gamma^h_{l\mu} - \sum_{\mu} R^h_{\mu jk} \Gamma^{\mu}_{li} - \sum_{\mu} R^h_{i\mu k} \Gamma^{\mu}_{lj} - \sum_{\mu} R^h_{ij\mu} \Gamma^{\mu}_{lk}.$$

From these equations we see that the left side of (3) equals

$$\left(\frac{\partial R^h_{ijk}}{\partial x^l} + \sum_{\mu} R^{\mu}_{ijk} \Gamma^h_{l\mu} - \sum_{\mu} R^h_{\mu jk} \Gamma^{\mu}_{li} \right) + \text{the two terms obtained by cyclically permuting } j, k, l.$$

Plugging back into the definition, we find that this is zero. ♦

Believe it or not, the Bianchi identity will be useful later on, and crucial in the last chapter of Volume V. Even more surprising, in Chapter 7 we will present a derivation of the Bianchi identity which will make it seem like a natural result. With the present proof of the Bianchi identity we end our summary of the classical theory of connections. The presentation was made mainly as background for the succeeding chapters, in which the same results will begin to take on a more modern appearance.

CHAPTER 6

THE ∇ OPERATOR

The contents of this chapter really differ very little from those of the previous one, but everything will look quite different. The clean modern symbolism which gets introduced here is abandoned only in those parts of the chapter which compare the present treatment with that given previously. This refurbishment of the classical theory, due to Koszul, is effected by singling out for invariant treatment just one of the concepts introduced previously, and then defining the other concepts in terms of it. We will begin with a definition, and then compare it to the classical one.

A **(Koszul) connection** on a C^∞ manifold M is a function ∇ (read “dell”) which associates a C^∞ vector field $\nabla_X Y$ to any two C^∞ vector fields X and Y , and which satisfies

- (1) $\nabla_{X_1+X_2} Y = \nabla_{X_1} Y + \nabla_{X_2} Y$
- (2) $\nabla_X (Y_1 + Y_2) = \nabla_X Y_1 + \nabla_X Y_2$
- (3) $\nabla_{fX} Y = f \cdot \nabla_X Y$
- (4) $\nabla_X (fY) = f \cdot \nabla_X Y + X(f) \cdot Y.$

Notice that ∇ is assumed linear *over the C^∞ functions* in the argument X . By our standard theorem (I.4-2) this means that for any given vector field Y we can define

$$\nabla_{X_p} Y \in M_p$$

for every $X_p \in M_p$ so that

$$\nabla_X Y = p \mapsto \nabla_{X_p} Y.$$

Alternatively, one can define a Koszul connection to be a map ∇ which assigns a vector $\nabla_{X_p} Y \in M_p$ to every $X_p \in M_p$ and every C^∞ vector field Y , and which satisfies

- (1') $\nabla_{X_p+X'_p} Y = \nabla_{X_p} Y + \nabla_{X'_p} Y$
- (2') $\nabla_{X_p} (Y_1 + Y_2) = \nabla_{X_p} Y_1 + \nabla_{X_p} Y_2$

$$(3') \quad \nabla_{aX_p} Y = a \nabla_{X_p} Y \text{ for all } a \in \mathbb{R}$$

$$(4') \quad \nabla_{X_p}(fY) = f(p) \cdot \nabla_{X_p} Y + X_p(f) \cdot Y_p$$

$$(5') \quad \text{If } X \text{ and } Y \text{ are } C^\infty \text{ vector fields, then so is } p \mapsto \nabla_{X_p} Y.$$

We then define $\nabla_X Y$ by $\nabla_X Y(p) = \nabla_{X_p} Y$.

As one example of a Koszul connection, we take M to be \mathbb{R}^n and let $\nabla_{X_p} Y$ be the directional derivative of Y in the direction X_p (computed by taking the directional derivative of the component functions of Y). It is clear that properties (1')–(5') hold for this ∇ . However, the most important justification for the particular conditions required of a Koszul connection comes from a comparison with classical connections. If x^1, \dots, x^n is a coordinate system on M , and we define Γ_{ij}^k by

$$(*) \quad \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \sum_{k=1}^n \Gamma_{ij}^k \frac{\partial}{\partial x^k},$$

then from (1)–(4) it is an easy exercise to deduce that Γ_{ij}^k are the components of a (classical) connection; conversely, given a classical connection, we can use (*) and (1)–(4) to determine a well-defined ∇ . In the coordinate system x , if

$$X = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i}, \quad Y = \sum_{j=1}^n \lambda^j \frac{\partial}{\partial x^j},$$

then

$$\nabla_X Y = \sum_{k=1}^n \left(\sum_{i=1}^n a^i \lambda^k_{;i} \right) \frac{\partial}{\partial x^k}.$$

For a given vector field Y we have a tensor ∇Y of type $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, that is, a collection of linear transformations $\nabla Y(p): M_p \rightarrow M_p$, given by

$$\nabla Y(p)(X_p) = \nabla_{X_p} Y \quad (= \nabla_X Y(p)).$$

Clearly

$$\nabla Y \left(\frac{\partial}{\partial x^i} \right) = \sum_{k=1}^n \lambda^k_{;i} \frac{\partial}{\partial x^k};$$

we can also write this as

$$\nabla Y = \sum_{i,k=1}^n \lambda^k_{;i} dx^i \otimes \frac{\partial}{\partial x^k},$$

which shows that ∇Y is the tensor classically described in terms of its components $\lambda^k_{;i}$.

The covariant derivatives of all other tensors are now going to be defined in terms of this covariant derivative, which we have made the cornerstone of our new definition of a connection (and *not* merely in terms of the Γ^k_{ij} which it determines). There are two completely different ways of doing this, each of which has its advantages. The first way is purely formal:

1. PROPOSITION. Let X be a C^∞ vector field on a C^∞ manifold M with a Koszul connection ∇ . Then there is a unique operator

$$A \mapsto \nabla_X A$$

from C^∞ tensor fields to C^∞ tensor fields, preserving the type $\binom{k}{l}$, such that

- (1) $\nabla_X f = X(f)$
- (2) $\nabla_X Y$ is the vector field given by the connection ∇
- (3) $A \mapsto \nabla_X A$ is linear over \mathbb{R}
- (4) $\nabla_X(A \otimes B) = \nabla_X A \otimes B + A \otimes \nabla_X B$
- (5) For any contraction C , we have $\nabla_X \circ C = C \circ \nabla_X$.

Moreover, each $\nabla_X A$ is linear *over the C^∞ functions* in the argument X , so for every tensor field A of type $\binom{k}{l}$ and every $X_p \in M_p$ we can define

$$\nabla_{X_p} A \in \mathcal{T}_l^k(M_p)$$

with

$$\begin{aligned} \nabla_{X_p + X'_p} A &= \nabla_{X_p} A + \nabla_{X'_p} A \\ \nabla_{aX_p} A &= a \nabla_{X_p} A. \end{aligned}$$

PROOF. Essentially, this is Problem I.5-15. If we define

$$\begin{aligned} Df &= X(f) \\ DY &= \nabla_X Y, \end{aligned}$$

then D does satisfy the condition

$$D(fY) = fDY + Df \cdot Y \quad \text{i.e.,} \quad D(f \otimes Y) = f \otimes DY + Df \otimes Y$$

which is assumed for this problem. Briefly, the proof that $D = \nabla$ can be extended uniquely is as follows. For a 1-form ω we want

$$\begin{aligned} X(\omega(Y)) &= D(\omega(Y)) = D(\text{contraction of } \omega \otimes Y) \\ &= \text{contraction of } [D\omega \otimes Y + \omega \otimes DY] \\ &= D\omega(Y) + \omega(DY) \\ &= D\omega(Y) + \omega(\nabla_X Y), \end{aligned}$$

so we want

$$(\nabla_X \omega)(Y) = D\omega(Y) = X(\omega(Y)) - \omega(\nabla_X Y) \quad \text{for all } Y.$$

Since any A is a sum of functions times tensor products of vector fields and 1-forms, conditions (3) and (4) determine $\nabla_X A$.

Following through the proof in detail, it is easily checked that $\nabla_X A$ is linear over the C^∞ functions in the argument X . ♦

In view of Proposition 1, for any tensor A of type $\binom{k}{l}$ we can define a new tensor ∇A of type $\binom{k+1}{l}$ by

$$\nabla A(p)(X_{1p}, \dots, X_{kp}, X_p) = \nabla_{X_p} A(X_{1p}, \dots, X_{kp}).$$

If $\omega = \sum_j \lambda_j dx^j$ is a tensor of type $\binom{1}{0}$, and we set

$$\nabla_{\frac{\partial}{\partial x^i}} \omega = \sum_{l=1}^n a_l dx^l,$$

then we have

$$\begin{aligned} \frac{\partial}{\partial x^i} \lambda_k &= \nabla_{\frac{\partial}{\partial x^i}} \left(\text{contraction of } \omega \otimes \frac{\partial}{\partial x^k} \right) \\ &= \text{contraction of } \left(\left[\nabla_{\frac{\partial}{\partial x^i}} \omega \right] \otimes \frac{\partial}{\partial x^k} + \omega \otimes \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^k} \right) \\ &= a_k + \sum_{\mu=1}^n \Gamma_{ik}^\mu \lambda_\mu, \end{aligned}$$

so we obtain

$$\begin{aligned} \nabla_{\frac{\partial}{\partial x^i}} \omega &= \sum_{k=1}^n \left(\frac{\partial \lambda_k}{\partial x^i} - \sum_{\mu=1}^n \Gamma_{ik}^\mu \lambda_\mu \right) dx^k \\ &= \sum_{k=1}^n \lambda_{k;i} dx^k. \end{aligned}$$

We can also write this as

$$\nabla \omega = \sum_{i,k=1}^n \lambda_{k;i} dx^k \otimes dx^i,$$

which shows that $\nabla\omega$ is the tensor classically described in terms of its components $\lambda_{k,i}$. Similarly, we easily see that if

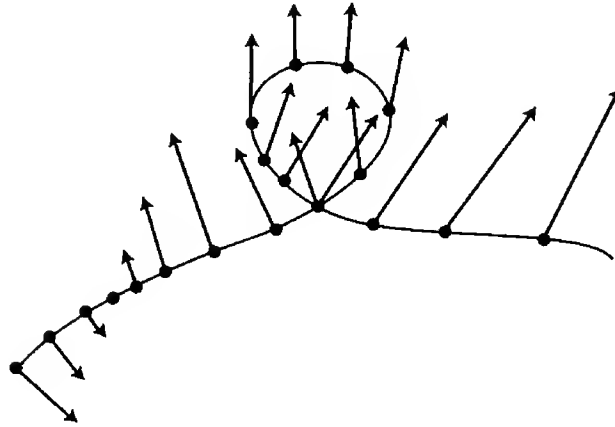
$$A = \sum A_{i_1 \dots i_k}^{j_1 \dots j_l} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}},$$

then

$$\nabla A = \sum A_{i_1 \dots i_k; h}^{j_1 \dots j_l} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes dx^h \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}}.$$

The uniqueness clause in Proposition 1 is precisely what accounts for its usefulness: all properties of ∇A should be derivable from properties (1)–(5), since these properties characterize ∇A . On the other hand, the proposition gives no idea what is going on geometrically. To obtain such a picture we introduce another extremely important concept.

First consider a curve $c: [a, b] \rightarrow M$. By a **vector field V along c** we mean a function V on $[a, b]$ with $V_t = V(t) \in M_{c(t)}$. In a coordinate system (x, U) we



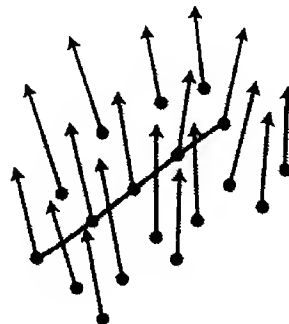
can write

$$V_t = \sum_{i=1}^n v_i(t) \cdot \frac{\partial}{\partial x^i} \Big|_{c(t)}.$$

We call V a C^∞ vector field along c if the functions v^i are C^∞ on $[a, b]$; this is equivalent to saying that $t \mapsto V_t(f)$ is C^∞ for every C^∞ function f on M .

Now suppose that V is a C^∞ vector field on a neighborhood of $c([a, b])$. Then

$$t \mapsto \nabla_{\frac{dc}{dt}} V$$



is a C^∞ vector field along c . This vector field is called the **covariant derivative of V along c** ; we will denote it by the convenient symbolism

$$\frac{DV}{dt},$$

which involves all the classical ambiguities. We would like to generalize this covariant derivative along c to vector fields V which are themselves defined only along c .

2. PROPOSITION. There is precisely one operation $V \mapsto DV/dt$, from C^∞ vector fields V along c to C^∞ vector fields along c , with the following properties:

- (a) $\frac{D(V+W)}{dt} = \frac{DV}{dt} + \frac{DW}{dt}$
- (b) $\frac{D(fV)}{dt} = \frac{df}{dt}V + f\frac{DV}{dt}$ for $C^\infty f: [a, b] \rightarrow \mathbb{R}$
- (c) If $V_s = Y_{c(s)}$ for some C^∞ vector field Y defined in a neighborhood of $c(t)$, then

$$\frac{DV}{dt} = \nabla_{\frac{dc}{dt}} Y.$$

PROOF. If x is a coordinate system around $p = c(t_0)$ then for t sufficiently close to t_0 we can write

$$V(t) = \sum_{j=1}^n v^j(t) \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)}$$

for unique functions v^j . If (a), (b), (c) are to hold, then we must have

$$\frac{DV}{dt} = \sum_{j=1}^n \frac{D}{dt} \left(v^j(t) \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)} \right) \quad \text{by (a)}$$

$$= \sum_{j=1}^n \left\{ \frac{dv^j}{dt} \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)} + v^j(t) \frac{D}{dt} \frac{\partial}{\partial x^j} \Big|_{c(t)} \right\} \quad \text{by (b)}$$

$$= \sum_{j=1}^n \left\{ \frac{dv^j}{dt} \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)} + v^j(t) \nabla_{\frac{dc}{dt}} \frac{\partial}{\partial x^j} \right\} \quad \text{by (c)}$$

$$= \sum_{j=1}^n \left\{ \frac{dv^j}{dt} \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)} + v^j(t) \sum_{i=1}^n \frac{dc^i}{dt} \nabla_{\frac{\partial}{\partial x^i}} \Big|_{c(t)} \frac{\partial}{\partial x^j} \right\}$$

and thus

$$\frac{DV}{dt} = \sum_{k=1}^n \left(\frac{dv^k}{dt} + \sum_{i,j=1}^n \Gamma_{ij}^k(c(t)) \frac{dc^i}{dt} v^j(t) \right) \frac{\partial}{\partial x^k} \Big|_{c(t)}$$

So there is at most one such operation. Conversely, it is easy to check that this formula does have the required properties. \blacklozenge

Remark. There are two things which should be noted about this Proposition, and the interest of the first tends to overshadow the significance of the second.

(1) The Proposition assigns a value to DV/dt even at points where $dc/dt = 0$. This value is not necessarily 0. In fact, if c is a constant curve, $c(t) = p$ for all t , then a vector field V along c is just a curve in M_p , and DV/dt is just the ordinary derivative of this vector-valued curve.

(2) When $dc/dt \neq 0$, so that c is an imbedding in a neighborhood of t , we can always write $V_s = Y_{c(s)}$ for some Y defined in a neighborhood of $c(t)$. But Y is not unique, so even now condition (c) does not by itself determine DV/dt —we also need conditions (a) and (b). This result is very similar to our basic principle for defining tensors (Theorem I.4-2): in the proof of that result we expressed all vectors in terms of the $\partial/\partial x^i$; in the present case we express all vector fields along c in terms of the $\partial/\partial x^i|_{c(t)}$.

We say that a vector field V along c is **parallel along c** (with respect to ∇) if $DV/dt = 0$ along c . When $M = \mathbb{R}^n$ and ∇ is just the directional derivative, we obtain the standard picture of a parallel vector field. In general, given a



curve $c: [a, b] \rightarrow M$, and a vector $V_a \in M_{c(a)}$, there is a unique vector field V along c which is parallel along c . This is because the equations

$$(*) \quad \frac{dv^k(t)}{dt} + \sum_{i,j=1}^n \frac{dc^i(t)}{dt} \Gamma_{ij}^k(c(t)) v^j(t) = 0$$

are linear differential equations with unique solutions v^j , defined on all of $[a, b]$, for given initial conditions; the desired vector field V is then

$$V_t = \sum_{j=1}^n v^j(t) \cdot \frac{\partial}{\partial x^j} \Big|_{c(t)}.$$

The vector $V_t \in M_{c(t)}$ is said to be obtained from V_a by **parallel translation along c** . It is clear from equations (*) that $(V+W)_t = V_t + W_t$ and $(\lambda \cdot V)_t = \lambda \cdot V_t$ for $\lambda \in \mathbb{R}$. We therefore obtain a linear transformation

$$\tau_t : M_{c(a)} \rightarrow M_{c(t)} \quad V_a \mapsto V_t.$$

Clearly, τ_t is one-one, for its inverse is just parallel translation along the reversed portion of c from t to a . Thus, along any curve c we obtain an isomorphism between any two tangent spaces $M_{c(t_1)}$ and $M_{c(t_2)}$; this possibility of comparing, or “connecting”, tangent spaces at different points gives rise to the term “connection”. It was invented by Levi-Civita, who used the equations (*) as the definition.

The parallel translation τ_t is defined in terms of ∇ , but we can also reverse the process.

3. PROPOSITION. Let c be a curve with $c(0) = p$ and $c'(0) = X_p$. Then

$$\nabla_{X_p} Y = \lim_{h \rightarrow 0} \frac{1}{h} (\tau_h^{-1} Y_{c(h)} - Y_p).$$

PROOF. Let V_1, \dots, V_n be parallel vector fields along c which are linearly independent at $c(0)$, and hence at all points of c . Set

$$Y(c(t)) = \sum_{i=1}^n \gamma^i(t) \cdot V_i(t).$$

Then

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{1}{h} (\tau_h^{-1} Y_{c(h)} - Y_p) &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\sum_{i=1}^n \gamma^i(h) \tau_h^{-1} V_i(h) - \gamma^i(0) V_i(0) \right] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left[\sum_{i=1}^n \gamma^i(h) V_i(0) - \gamma^i(0) V_i(0) \right] \\ &= \sum_{i=1}^n \lim_{h \rightarrow 0} \frac{\gamma^i(h) - \gamma^i(0)}{h} \cdot V_i(0) \\ &= \sum_{i=1}^n \frac{d\gamma^i}{dt}(0) V_i(0) = \frac{D}{dt} \Big|_{t=0} \sum_{i=1}^n \gamma^i(t) V_i(t) \\ &= \nabla_{X_p} Y. \quad \spadesuit \end{aligned}$$

Motivated by Proposition 3, we now define $\nabla_{X_p} A$ where A is any tensor field of type $\binom{k}{l}$. We have

$$A(q) \in \mathcal{T}_l^k(M_q) \quad \text{for all } q,$$

and the isomorphism

$$\tau_t: M_{c(0)} \rightarrow M_{c(t)}$$

gives rise to an isomorphism

$$\mathcal{T}_l^k(\tau_t): \mathcal{T}_l^k(M_{c(0)}) \rightarrow \mathcal{T}_l^k(M_{c(t)});$$

therefore we can define

$$\nabla_{X_p} A = \lim_{h \rightarrow 0} \frac{1}{h} ([\mathcal{T}_l^k(\tau_h)]^{-1} A(c(h)) - A(p)).$$

If we regard $A(q)$ as a function of k tangent vectors in M_q and l vectors in M_q^* , this means that for $v_1, \dots, v_k \in M_p$ and $\lambda_1, \dots, \lambda_l \in M_p^*$ we have

$$\begin{aligned} (\nabla_{X_p} A)(v_1, \dots, v_k, \lambda_1, \dots, \lambda_l) \\ = \lim_{h \rightarrow 0} \frac{1}{h} [A(c(h))(\tau_h v_1, \dots, \tau_h v_k, \tau_h^* \lambda_1, \dots, \tau_h^* \lambda_l) \\ - A(p)(v_1, \dots, v_k, \lambda_1, \dots, \lambda_l)]. \end{aligned}$$

4. PROPOSITION. This $\nabla_{X_p} A$ coincides with that given by Proposition 1.

PROOF. It suffices to prove properties (1)–(5) for the new $\nabla_{X_p} A$. This is left to the reader [the proof of (4) and (5) involves the usual trick which one uses in the proof of the product rule for derivatives]. ♦

[Note that without this result it would not be at all obvious that $\nabla_{X_p} A$ is linear in X_p .]

5. COROLLARY. Let A be a tensor field of type $\binom{k}{l}$. If Y_1, \dots, Y_k are vector fields, then

$$(\nabla_{X_p} A)(Y_{1p}, \dots, Y_{kp}) = \nabla_{X_p} (A(Y_1, \dots, Y_k)) - \sum_{i=1}^k A(Y_{1p}, \dots, \nabla_{X_p} Y_i, \dots, Y_{kp}).$$

PROOF. Use Proposition 1, noting that $A(Y_1, \dots, Y_k)$ can be obtained by applying k contractions to

$$A \otimes Y_1 \otimes \dots \otimes Y_k.$$

(It is also instructive to obtain a proof from the definition in terms of parallel translations.) ♦

Our definition of $\nabla_X A$ may be compared to the definition, in Problem I.5-14, of $L_X A$. In the latter case, the maps ϕ^*_h given by the vector field X play the roles of the parallel translations τ_h in the present instance. Notice that Problem I.5-15 can be used in both cases to formally extend the operation from vector fields to arbitrary tensors.

After these preliminaries, the further study of Koszul connections proceeds quite rapidly. We first define, for vector fields X and Y ,

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y].$$

A simple calculation shows that T is *linear over the C^∞ functions*, so that it determines a tensor. Clearly this is just the classical torsion tensor; as usual, the invariant definition involves a bracket, which disappears in the expression in coordinates.

We now want to distinguish the connection determined by the Christoffel symbols for a metric. Suppose we are in a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$. We will call a connection **compatible with $\langle \cdot, \cdot \rangle$** if the parallel translations $\tau_t: M_{c(a)} \rightarrow M_{c(t)}$ along any curve $c: [a, b] \rightarrow M$ are isometries (with respect to $\langle \cdot, \cdot \rangle_{c(a)}$ and $\langle \cdot, \cdot \rangle_{c(t)}$).

6. LEMMA. A connection ∇ is compatible with a metric $\langle \cdot, \cdot \rangle$ if and only if it satisfies the following condition: If V and W are vector fields along any curve c , then

$$\frac{d}{dt} \langle V, W \rangle = \left\langle \frac{DV}{dt}, W \right\rangle + \left\langle V, \frac{DW}{dt} \right\rangle.$$

PROOF. Suppose ∇ satisfies this condition. Then if V is parallel along c we have

$$\frac{d}{dt} \langle V, V \rangle = 2 \left\langle \frac{DV}{dt}, V \right\rangle = 0,$$

so $\langle V, V \rangle$ is constant along c . Thus each τ_t is norm preserving, and hence an isometry.

Conversely, suppose ∇ is compatible with the metric. Choose parallel vector fields P_1, \dots, P_n along c which are orthonormal at one point of c , and hence at every point of c . Let

$$V_t = \sum_{i=1}^n v^i(t) P_{i_t}, \quad W_t = \sum_{j=1}^n w^j(t) P_{j_t}.$$

Then

$$\langle V, W \rangle = \sum_{i=1}^n v^i \cdot w^i,$$

and by Proposition 2 we have, remembering that $DP_i/dt = 0$,

$$\frac{DV}{dt} = \sum_{i=1}^n \frac{dv^i}{dt} P_i, \quad \frac{DW}{dt} = \sum_{j=1}^n \frac{dw^j}{dt} P_j.$$

So

$$\left\langle \frac{DV}{dt}, W \right\rangle + \left\langle V, \frac{DW}{dt} \right\rangle = \sum_{i=1}^n \left(\frac{dv^i}{dt} w^i + v^i \frac{dw^i}{dt} \right) = \frac{d}{dt} \langle V, W \rangle. \quad \spadesuit$$

7. COROLLARY. The connection ∇ is compatible with $\langle \cdot, \cdot \rangle$ if and only if

$$X_p \langle Y, Z \rangle = \langle \nabla_{X_p} Y, Z_p \rangle + \langle Y_p, \nabla_{X_p} Z \rangle$$

for all vector fields Y, Z and vectors $X_p \in M_p$.

PROOF. Apply the Lemma to a curve c with $c'(0) = X_p$. \spadesuit

8. LEMMA (FUNDAMENTAL LEMMA OF RIEMANNIAN GEOMETRY). On a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ there is a unique symmetric connection compatible with $\langle \cdot, \cdot \rangle$.

PROOF. Suppose ∇ is compatible with $\langle \cdot, \cdot \rangle$. Choose a coordinate system (x, U) . By Corollary 7

$$(*) \quad \frac{\partial g_{jk}}{\partial x^i} = \frac{\partial}{\partial x^i} \left\langle \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k} \right\rangle = \left\langle \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k} \right\rangle + \left\langle \frac{\partial}{\partial x^j}, \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^k} \right\rangle.$$

Cyclically permuting i, j, k , and using

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i} \quad (\text{from symmetry}),$$

we obtain

$$\sum_{l=1}^n \Gamma_{ij}^l g_{lk} = \left\langle \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k} \right\rangle = [ij, k],$$

which implies that

$$\Gamma_{ij}^l = \sum_{k=1}^n g^{kl} [ij, k].$$

Thus the Γ 's for ∇ must be the Christoffel symbols.

We know that the Christoffel symbols do indeed satisfy $(*)$, which shows that the equation in Corollary 7 does hold. (In fact, this equation is equivalent to Ricci's Lemma, 5-3.) \spadesuit

The unique connection of Lemma 8 obviously ought to be called the Christoffel connection for $\langle \cdot, \cdot \rangle$; instead, it is called the Levi-Civita connection* for $\langle \cdot, \cdot \rangle$! Naturally, Lemma 8 is more impressive before reading the proof than after. Nevertheless, it is still a very nice result. Perhaps its only defect is the restriction to *symmetric* connections; in Addendum 1 to this chapter we present some justification for this restriction. We have already given one interpretation of symmetry; but the following will be more useful for present purposes. For a C^∞ function $s: \mathbb{R}^2 \rightarrow M$ (a “parameterized surface” in M), we define a **vector field V along s** to be a function V with

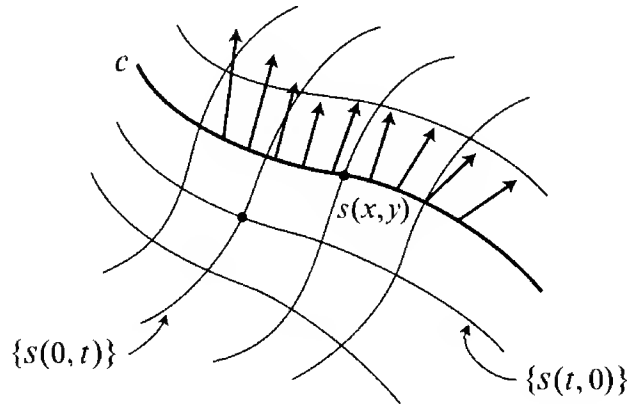
$$V_{(x,y)} \in M_{s(x,y)}.$$

In particular, we have the vector fields

$$\frac{\partial s}{\partial x} = s_* \left(\frac{\partial}{\partial x} \right), \quad \frac{\partial s}{\partial y} = s_* \left(\frac{\partial}{\partial y} \right).$$

For any C^∞ vector field V along s , we define

$$\left(\frac{DV}{\partial x} \right)_{(x,y)} = \begin{array}{l} \text{covariant derivative along } c(t) = s(t, y) \\ \text{of } t \mapsto V_{(t,y), \text{ evaluated at } t = x,} \end{array}$$



and we define $DV/\partial y$ similarly.

8. PROPOSITION. If ∇ is symmetric, then

$$\frac{D}{\partial x} \frac{\partial s}{\partial y} = \frac{D}{\partial y} \frac{\partial s}{\partial x}.$$

PROOF. Express both sides in terms of a coordinate system and compute. ♦

*This is an historical mix-up, due to the fact that the collection of parallel translations determined by ∇ was called “Levi-Civita’s connection” for ∇ .

We are now ready to define the curvature tensor for a Koszul connection ∇ . If X, Y, Z are vector fields, we consider the vector field

$$R(X, Y)Z = \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X, Y]}Z.$$

A straightforward computation shows that (because of the bracket term) R is *linear over the C^∞ functions* in all three variables, so it defines a tensor R , the **curvature tensor** of ∇ . Obviously, this definition is somehow related to the Ricci identity (Proposition 5-8)

$$(*) \quad \lambda^i{}_{;jk} - \lambda^i{}_{;kj} = - \sum_{l=1}^n \lambda^l R^i{}_{ljk} + \sum_{l=1}^n \lambda^i{}_{;l} T^l{}_{jk},$$

but at first sight we seem to be missing a term for the torsion. To see why this is so, recall that if

$$Z = \sum_{i=1}^n \lambda^i \frac{\partial}{\partial x^i},$$

then $\lambda^i{}_{;j}$ are the components of ∇Z , i.e.,

$$\nabla_{\frac{\partial}{\partial x^j}} Z = \sum_{i=1}^n \lambda^i{}_{;j} \frac{\partial}{\partial x^i}.$$

Now $\lambda^i{}_{;jk} = \lambda^i{}_{;j;k}$ are *not* the components of $\nabla_{\partial/\partial x^k}(\nabla_{\partial/\partial x^j} Z)$; rather they are the components of $\nabla(\nabla Z)$. So

$$\begin{aligned} \sum_{i=1}^n \lambda^i{}_{;jk} \frac{\partial}{\partial x^i} &= \nabla \nabla Z \left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k} \right) \\ &= \left[\nabla_{\frac{\partial}{\partial x^k}} (\nabla Z) \right] \left(\frac{\partial}{\partial x^j} \right) \\ &= \nabla_{\frac{\partial}{\partial x^k}} \left(\nabla Z \left(\frac{\partial}{\partial x^j} \right) \right) - \nabla Z \left(\nabla_{\frac{\partial}{\partial x^k}} \frac{\partial}{\partial x^j} \right) \quad \text{by Corollary 5} \\ &= \nabla_{\frac{\partial}{\partial x^k}} \left(\nabla_{\frac{\partial}{\partial x^j}} Z \right) - \nabla \left(\nabla_{\frac{\partial}{\partial x^k}} \frac{\partial}{\partial x^j} \right) Z. \end{aligned}$$

Consequently,

$$\begin{aligned}
& \nabla_{\frac{\partial}{\partial x^j}} \left(\nabla_{\frac{\partial}{\partial x^k}} Z \right) - \nabla_{\frac{\partial}{\partial x^k}} \left(\nabla_{\frac{\partial}{\partial x^j}} Z \right) \\
&= \sum_{i=1}^n (\lambda^i{}_{;kj} - \lambda^i{}_{;jk}) \frac{\partial}{\partial x^i} + \nabla_{\left(\nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^k} - \nabla_{\frac{\partial}{\partial x^k}} \frac{\partial}{\partial x^j} \right)} Z \\
&= \sum_{i=1}^n \left(\sum_{l=1}^n \lambda^l R^i{}_{ljk} - \sum_{l=1}^n \lambda^i{}_{;l} T^l_{jk} \right) \frac{\partial}{\partial x^i} + \nabla_{\left(\sum_{l=1}^n T^l_{jk} \frac{\partial}{\partial x^l} \right)} Z \quad \text{by } (*).
\end{aligned}$$

Since

$$\nabla_{\left(\sum_{l=1}^n T^l_{jk} \frac{\partial}{\partial x^l} \right)} Z = \sum_{l=1}^n T^l_{jk} \nabla_{\frac{\partial}{\partial x^l}} Z = \sum_{i,l=1}^n T^l_{jk} \lambda^i{}_{;l} \frac{\partial}{\partial x^i},$$

we obtain simply

$$\begin{aligned}
\nabla_{\frac{\partial}{\partial x^j}} \left(\nabla_{\frac{\partial}{\partial x^k}} Z \right) - \nabla_{\frac{\partial}{\partial x^k}} \left(\nabla_{\frac{\partial}{\partial x^j}} Z \right) &= \sum_{i=1}^n \left(\sum_{l=1}^n \lambda^l R^i{}_{ljk} \right) \frac{\partial}{\partial x^i} \\
&= R \left(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k} \right) Z,
\end{aligned}$$

so the definition does agree with the classical one. At the same time, it is clearly preferable, in that it does not involve the torsion.

[Classically, there is practically no way to even name the quantity

$$\nabla_{\frac{\partial}{\partial x^j}} \left(\nabla_{\frac{\partial}{\partial x^k}} Z \right),$$

since $\lambda^i{}_{;kj}$ is automatically interpreted as the components of the tensor $\nabla \nabla Z$. For a vector field $Y = \sum_i b^i \partial / \partial x^i$, we can write the components of

$$\nabla_{\frac{\partial}{\partial x^j}} (\nabla_Y Z) \quad \text{as} \quad \left(\sum_{l=1}^n \lambda^i{}_{;l} b^l \right)_{;j},$$

the summation over l making it clear that we are taking covariant derivatives of the tensor field with components $\mu^i = \sum_l \lambda^i{}_{;l} b^l$. However, for the special case

$$\nabla_{\frac{\partial}{\partial x^j}} \left(\nabla_{\frac{\partial}{\partial x^k}} Z \right),$$

the expression

$$\left(\sum_{l=1}^n \lambda^i{}_{;l} \delta^l_k \right)_{;j}$$

is about the best we can do.]

The following result will be our analogue of the Ricci identities.

9. **PROPOSITION.** Let $s: \mathbb{R}^2 \rightarrow M$ be a parameterized surface, and V a C^∞ vector field along s . Then

$$\frac{D}{\partial x} \frac{D}{\partial y} V - \frac{D}{\partial y} \frac{D}{\partial x} V = R \left(\frac{\partial s}{\partial x}, \frac{\partial s}{\partial y} \right) V.$$

PROOF. Compute in a coordinate system. \diamond

It should come as no surprise to learn that we can now prove the Test Case; it may be surprising, however, to see how simple the proof becomes.

10. **THEOREM (THE TEST CASE; THIRD VERSION).** Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R (for the Levi-Civita connection) is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. We assume we are in \mathbb{R}^n , with y^1, \dots, y^n the standard coordinate system.

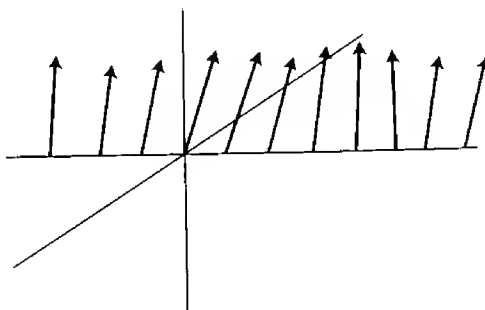
Step 1. We claim that we can find vector fields X , with arbitrary values $X(0) \in \mathbb{R}^n_0$, satisfying

$$\nabla_{\frac{\partial}{\partial y^i}} X = 0 \quad \text{for all } i,$$

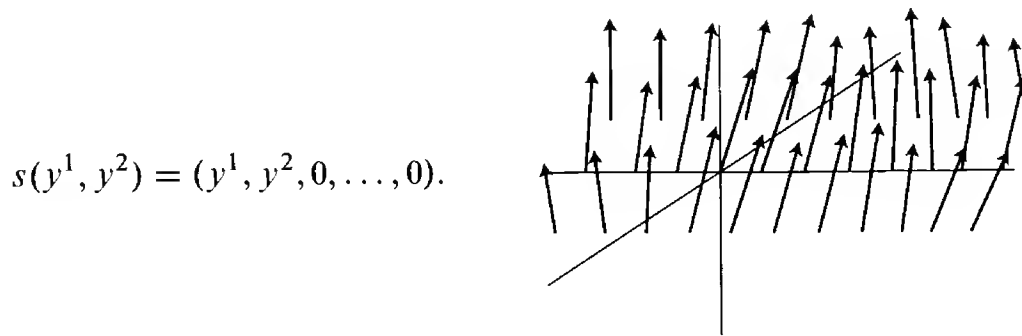
and hence

$$\nabla_Z X = 0 \quad \text{for all } Z.$$

To do this we first choose $X(y, 0, \dots, 0)$, with the desired initial value, so that it is parallel along the y^1 -axis.



For each fixed y^1 , we then choose $X(y^1, y, 0, \dots, 0)$, with the values $X(y^1, 0, \dots, 0)$ just obtained, so that X is parallel along the curves $y \mapsto (y^1, y, 0, \dots, 0)$. The vector field X is now defined on the surface



Clearly $DX/\partial y^2$ is 0 along s , while $DX/\partial y^1$ is 0 along $\{s(y, 0)\}$. Now we have

$$\frac{D}{\partial y^1} \frac{D}{\partial y^2} X - \frac{D}{\partial y^2} \frac{D}{\partial y^1} X = R \left(\frac{\partial s}{\partial y^1}, \frac{\partial s}{\partial y^2} \right) X = 0,$$

so

$$\frac{D}{\partial y^2} \frac{D}{\partial y^1} X = 0.$$

This means that $DX/\partial y^1$ is parallel along the curves $y \mapsto s(y^1, y)$. Since $DX/\partial y^1$ is 0 at $s(y^1, 0)$, we have $DX/\partial y^1 = 0$ along s .

We can clearly continue in this way to obtain the desired X . Now choose X_1, \dots, X_n with this property so that $X_1(0), \dots, X_n(0)$ are orthonormal with respect to $\langle \cdot, \cdot \rangle_0$. Clearly, X_1, \dots, X_n are linearly independent in a neighborhood of 0.

Step 2. Since the connection ∇ associated with $\langle \cdot, \cdot \rangle$ is symmetric, we have

$$0 = \nabla_{X_i} X_j - \nabla_{X_j} X_i - [X_i, X_j].$$

But $\nabla_Z X_i = 0$ for all Z . So $[X_i, X_j] = 0$. This means that there is a coordinate system x^1, \dots, x^n with $X_i = \partial/\partial x^i$.

Step 3. We claim that x is the desired coordinate system, i.e., that the X_i are everywhere orthonormal. This is obvious, for they are orthonormal at 0 and parallel along any curve, and parallel translation preserves the inner product $\langle \cdot, \cdot \rangle$. ♦

This proof is completely analogous to the first two proofs of the Test Case, except that it is “dual” to them in the sense that we find the vector fields $\partial/\partial x^i$ instead of the 1-forms dx^i .

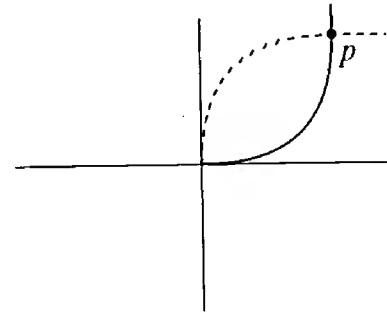
Step 1 uses the conditions $R = 0$ to obtain the parallel vector fields X_i , in a way completely analogous to, but simpler than, *Step 1* in the second version.

Step 2 now uses *symmetry of the connection* to prove that $[X_i, X_j] = 0$, a condition which is dual to exactness of the 1-forms obtained in the previous versions.

Step 3 uses the *definition of the Levi-Civita connection* to prove that the vectors X_i are everywhere orthonormal. As we have already pointed out, the fact that parallel translation preserves the inner product $\langle \cdot, \cdot \rangle$ is equivalent to Ricci’s Lemma, which we used in *Step 3* of the second version.

Notice that the proof shows that if we can find n everywhere linearly independent vector fields X_1, \dots, X_n , which are **parallel** (i.e., which satisfy $\nabla_Z X_i = 0$ for all Z), then the manifold is flat. Consequently, such vector fields generally *cannot* be found. This means that parallel translation of a vector along two different curves with the same end-points generally gives *different* results; for if X_p were the parallel translate of X_0 along both the curves shown below, we would clearly have both

$$\nabla_{\frac{\partial}{\partial x^1}} X(p) = 0 \quad \text{and} \quad \nabla_{\frac{\partial}{\partial x^2}} X(p) = 0.$$



This phenomenon can also be described by saying that parallel translation of a vector along a closed curve generally brings it back to a different vector. In Volume III we will see some more quantitative statements of this fact.

Our next two results are simply Proposition 5-9 and 4-10. They are reproved here in a spirit more in keeping with the present treatment of connections, but it is not hard to see that they are essentially the proofs given before. As a matter of notation, if we are given an expression $A(X, Y, Z)$, we will let $\mathfrak{S}A(X, Y, Z)$ denote the cyclic sum

$$\mathfrak{S}A(X, Y, Z) = A(X, Y, Z) + A(Y, Z, X) + A(Z, X, Y).$$

11. PROPOSITION. Let ∇ be a connection with torsion T and curvature R . Then for all vector fields X, Y, Z, W we have

$$(1) \quad R(X, Y)Z = -R(Y, X)Z$$

$$(2) \quad (\text{Bianchi's first identity})$$

$$\mathfrak{S}\{R(X, Y)Z\} = \mathfrak{S}\{(\nabla_X T)(Y, Z)\} + \mathfrak{S}\{T(T(X, Y)Z)\}$$

$$(3) \quad (\text{Bianchi's second identity})$$

$$\mathfrak{S}\{(\nabla_Z R)(X, Y, W)\} + \mathfrak{S}\{R(T(X, Y), Z)W\} = 0.$$

In particular, if $T = 0$, then

$$(2') \quad \mathfrak{S}\{R(X, Y)Z\} = 0$$

$$(3') \quad \mathfrak{S}\{(\nabla_Z R)(X, Y, W)\} = 0.$$

PROOF. (1) is clear from the definition.

To prove (2), we first note that

$$\begin{aligned} T(T(X, Y), Z) &= T(\nabla_X Y, Z) - T(\nabla_Y X, Z) - T([X, Y], Z) \\ &= T(\nabla_X Y, Z) + T(Z, \nabla_Y X) - T([X, Y], Z). \end{aligned}$$

We also have, by Corollary 5,

$$(\nabla_Z T)(X, Y) = \nabla_Z(T(X, Y)) - T(\nabla_Z X, Y) - T(X, \nabla_Z Y).$$

From these equations we obtain

$$\mathfrak{S}\{T(T(X, Y), Z)\} = -\mathfrak{S}\{(\nabla_Z T)(X, Y)\} + \mathfrak{S}\{\nabla_Z(T(X, Y)) - T([X, Y], Z)\}.$$

The second term on the right side equals

$$\mathfrak{S}\{\nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X, Y]}Z\} + \mathfrak{S}\{[[X, Y], Z]\} = \mathfrak{S}\{R(X, Y)Z\} + 0,$$

since the Jacobi identity states that $\mathfrak{S}\{[[X, Y], Z]\} = 0$.

To prove (3), we first use (1) to obtain

$$\begin{aligned} \mathfrak{S}\{R(T(X, Y), Z)W\} &= \mathfrak{S}\{R(\nabla_X Y, Z)W + R(Z, \nabla_Y X)W - R([X, Y], Z)W\} \\ &= \mathfrak{S}\{R(\nabla_Z X, Y)W + R(X, \nabla_Z Y)W\} - \mathfrak{S}\{R([X, Y], Z)W\}. \end{aligned}$$

By Corollary 5 we also have

$$\begin{aligned} (\nabla_Z R)(X, Y, W) &= \nabla_Z(R(X, Y)W) - R(X, Y)\nabla_Z W \\ &\quad - R(\nabla_Z X, Y)W - R(X, \nabla_Z Y)W. \end{aligned}$$

From these equations we have

$$\begin{aligned} \mathfrak{S}\{R(T(X, Y), Z)W\} &= -\mathfrak{S}\{(\nabla_Z R)(X, Y, W)\} \\ &\quad + \mathfrak{S}\{\nabla_Z(R(X, Y)W) - R(X, Y)\nabla_Z W \\ &\quad - R([X, Y], Z)W\}. \end{aligned}$$

Now

$$\begin{aligned} &\nabla_Z(R(X, Y)W) - R(X, Y)\nabla_Z W - R([X, Y], Z)W \\ &= \nabla_Z \nabla_X \nabla_Y W - \nabla_Z \nabla_Y \nabla_X W \quad \boxed{-\nabla_Z \nabla_{[X, Y]} W} \\ &\quad - \nabla_X \nabla_Y \nabla_Z W + \nabla_Y \nabla_X \nabla_Z W + \nabla_{[X, Y]} \nabla_Z W \\ &\quad \nabla_{[X, Y]} \nabla_Z W \quad + \quad \boxed{\nabla_Z \nabla_{[X, Y]} W} \\ &\quad + \nabla_{[[X, Y], Z]} W. \end{aligned}$$

Writing $[\nabla_X, \nabla_Y]$ for the operation $W \mapsto \nabla_X \nabla_Y W - \nabla_Y \nabla_X W$ (as on pg. I.155), we can write this as

$$\nabla_Z([\nabla_X, \nabla_Y]W) - [\nabla_X, \nabla_Y](\nabla_Z W) + \nabla_{[[X, Y], Z]} W.$$

The cyclic sum of these quantities is 0, because of the Jacobi identity

$$\mathfrak{S}\{[[X, Y], Z]\} = 0,$$

and the “Jacobi identity”

$$\mathfrak{S}\{[\nabla_Z, [\nabla_X, \nabla_Y]]W\} = 0$$

(recall that in any ring, if we define $[a, b] = ab - ba$, then $[,]$ satisfies the Jacobi identity). ♦

12. PROPOSITION. For the curvature tensor R of the Levi-Civita connection ∇ associated to a Riemannian metric \langle , \rangle we also have

- (1) $\langle R(X, Y)Z, W \rangle = -\langle R(X, Y)W, Z \rangle$
- (2) $\langle R(X, Y)Z, W \rangle = \langle R(Z, W)X, Y \rangle.$

PROOF. Equation (1) is equivalent to

$$\langle R(X, Y)Z, Z \rangle = 0 \quad \text{for all } X, Y, Z.$$

It suffices to prove this when $[X, Y] = 0$. In this case

$$\langle R(X, Y)Z, Z \rangle = \langle \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z), Z \rangle,$$

so we must show that $\langle \nabla_X(\nabla_Y Z), Z \rangle$ is symmetric in X and Y .

Now $YX\langle Z, Z \rangle$ is symmetric in X and Y , since $[X, Y] = 0$. But

$$X\langle Z, Z \rangle = 2\langle \nabla_X Z, Z \rangle,$$

so

$$YX\langle Z, Z \rangle = 2\langle \nabla_Y \nabla_X Z, Z \rangle + 2\langle \nabla_X Z, \nabla_Y Z \rangle.$$

Since the right-most term is symmetric in X and Y , so is $\langle \nabla_Y \nabla_X Z, Z \rangle$.

Equation (2) follows from (1), Proposition 12, and Proposition 4-11. ♦

To complete our treatment of Koszul connections, we define a **geodesic** for ∇ to be a path $\gamma: [a, b] \rightarrow M$ with

$$\frac{D}{dt} \frac{d\gamma}{dt} = 0;$$

thus, the tangent vector $d\gamma/dt$ must be parallel along γ . In the coordinate system x we immediately obtain the equations for a geodesic,

$$\frac{d^2 \gamma^k}{dt^2} + \sum_{i,j=1}^n \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt} = 0.$$

The existence of a unique geodesic, with given initial vector $\gamma'(0) \in M_p$ follows immediately. Since parallel translation is an isomorphism, the tangent vector $d\gamma/dt$ of a geodesic γ is nowhere zero (except when γ is a constant path). If V is any vector field along γ , then we can write

$$V = f \frac{d\gamma}{dt} \quad f: [a, b] \rightarrow \mathbb{R},$$

and

$$\frac{DV}{dt} = \frac{df}{dt} \frac{d\gamma}{dt} + f \frac{D}{dt} \frac{d\gamma}{dt} = \frac{df}{dt} \frac{d\gamma}{dt}.$$

This is 0 only if f is linear. Consequently, a reparameterization $\bar{\gamma} = \gamma \circ p$ of γ is also a geodesic only if p is linear.*

We know that this result can be made more precise for the Levi-Civita connection ∇ associated to a metric $\langle \cdot, \cdot \rangle$, and the proof of the more precise result is now especially easy: We have

$$\frac{d}{dt} \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = 2 \left\langle \frac{D}{dt} \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = 0,$$

so $\|d\gamma/dt\|$ is constant, i.e., γ is parameterized proportionally to arclength (Theorem I.9-12). The reader is invited to investigate how the proof of Gauss' Lemma (I.9-15) is simplified in the present set up. We will complete our study of the ∇ operator by providing the invariant description of the First Variation Formula promised so long ago.

13. THEOREM (FIRST VARIATION FORMULA). Let $\gamma: [a, b] \rightarrow M$ be a piecewise C^∞ path and $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ a variation. Let

$$\begin{aligned} W_t &= \frac{\partial \alpha}{\partial u}(0, t), & \text{the "variation vector field"} \\ V_t &= \frac{d\gamma}{dt}, & \text{the "velocity vector of } \gamma \text{"} \\ A_t &= \frac{D}{dt} V_t, & \text{the "acceleration vector of } \gamma \text{"}. \end{aligned}$$

Also choose $a = t_0 < \cdots < t_N = b$ to include all discontinuity points of V , and set

$$\begin{aligned} \Delta_{t_i} V &= V(t_i^+) - V(t_i^-) & i = 1, \dots, N-1 \\ \Delta_{t_0} V &= V(t_0^+) \\ \Delta_{t_N} V &= -V(t_N^-). \end{aligned}$$

Then

$$\left. \frac{dE(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_a^b \langle W_t, A_t \rangle dt - \sum_{i=0}^N \langle W_{t_i}, \Delta_{t_i} V \rangle.$$

PROOF. We will give the proof when V has no discontinuities, leaving to the reader the simple auxiliary argument for the general case.

*Again we point out that certain global theorems about geodesics for Riemannian metrics do not generalize to arbitrary connections; see Chapter 8, Addendum 2.

From

$$\frac{\partial}{\partial u} \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle = 2 \left\langle \frac{D}{\partial u} \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle$$

we obtain

$$\begin{aligned} \frac{dE(\bar{\alpha}(u))}{du} &= \frac{1}{2} \frac{d}{du} \int_a^b \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle dt = \int_a^b \left\langle \frac{D}{\partial u} \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle dt \\ &= \int_a^b \left\langle \frac{D}{\partial t} \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle dt \quad \text{by Proposition 9.} \end{aligned}$$

Now the identity

$$\frac{\partial}{\partial t} \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle = \left\langle \frac{D}{\partial t} \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle + \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle$$

implies the following analogue of integration by parts:

$$\int_a^b \left\langle \frac{D}{\partial t} \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle dt = \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle \Big|_{t=a}^{t=b} - \int_a^b \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt.$$

Thus

$$\frac{dE(\bar{\alpha}(u))}{du} \Big|_{u=0} = - \int_a^b \langle W_t, A_t \rangle dt - \langle W(a), V(a) \rangle + \langle W(b), V(b) \rangle,$$

which is the desired formula. ♦

ADDENDUM 1

CONNECTIONS WITH THE SAME GEODESICS

Let ∇ and $\bar{\nabla}$ be two connections on M . We define

$$D(X, Y) = \bar{\nabla}_X Y - \nabla_X Y.$$

A simple computation shows that D is linear over the C^∞ functions in both arguments, so it determines a tensor D , the **difference tensor** of the two connections. In a coordinate system x we have

$$D = \sum \bar{\Gamma}_{ij}^k - \Gamma_{ij}^k \, dx^i \otimes dx^j \otimes \frac{\partial}{\partial x^k}.$$

As is well known, there is a unique way to write $D = S + A$ with S symmetric and A alternating, namely

$$S(X, Y) = \frac{1}{2}[D(X, Y) + D(Y, X)]$$

$$A(X, Y) = \frac{1}{2}[D(X, Y) - D(Y, X)].$$

Note that if \bar{T} and T are the torsion tensors of $\bar{\nabla}$ and ∇ , then

$$\begin{aligned} (*) \quad 2A(X, Y) &= \bar{\nabla}_X Y - \nabla_X Y - \bar{\nabla}_Y X + \nabla_Y X \\ &= \bar{T}(X, Y) + [X, Y] - T(X, Y) - [X, Y] \\ &= \bar{T}(X, Y) - T(X, Y), \end{aligned}$$

so $\bar{\nabla}$ and ∇ have the same torsion if and only if $A = 0$.

14. PROPOSITION. The following are equivalent:

- (a) The connections $\bar{\nabla}$ and ∇ have the same geodesics [with the same parameterizations]
- (b) $D(X, X) = 0$ for all X
- (c) $S = 0$.

PROOF. (a) \Rightarrow (b): Given $0 \neq X_p \in M$, let γ be the geodesic (for the connections $\bar{\nabla}$ and ∇) with $\gamma'(0) = X_p$. Let X be a vector field in a neighborhood of p which equals $d\gamma/dt$ along the part of γ in the neighborhood. Then

$$D(X_p, X_p) = \bar{\nabla}_{X_p} X - \nabla_{X_p} X = 0 + 0.$$

(b) \Rightarrow (a): Let γ be a geodesic for ∇ , and let X be a vector field which equals $d\gamma/dt$ along a portion of γ . Then

$$\bar{\nabla}_{X_p} X = D(X_p, X_p) + \nabla_{X_p} X = 0 + 0,$$

which shows that γ is a geodesic for $\bar{\nabla}$.

(b) \Leftrightarrow (c): $D(X, X) = 0 \Leftrightarrow S(X, X) = 0$. The latter condition implies that $S = 0$, for we have

$$\begin{aligned} 0 &= S(X + Y, X + Y) = S(X, X) + S(Y, Y) + S(X, Y) + S(Y, X) \\ &= 2S(X, Y). \quad \spadesuit \end{aligned}$$

15. COROLLARY. If the connections ∇ and $\bar{\nabla}$ have the same geodesics and the same torsion, then $\bar{\nabla} = \nabla$.

16. COROLLARY. For every connection $\bar{\nabla}$, there is a unique connection ∇ with the same geodesics and with torsion 0.

PROOF. Uniqueness follows from Corollary 16. For existence we define $\nabla_X Y = \bar{\nabla}_X Y - \frac{1}{2}\bar{T}(X, Y)$, checking easily that ∇ is a connection. Since \bar{T} is skew-symmetric, $D = \frac{1}{2}\bar{T}$ must simply be A , so $S = 0$, so ∇ and $\bar{\nabla}$ have the same geodesics. Also, $T = \bar{T} - 2A = 0$, so ∇ has torsion 0. \spadesuit

We thus see that if we divide all connections into equivalence classes, putting all connections with the same geodesics in the same class, then each class has exactly one connection with zero torsion. We can also say exactly how large each class is: If ∇ is a connection with torsion 0 and \bar{T} is a skew-symmetric tensor of type $\binom{2}{0}$, then there is a connection $\bar{\nabla}$ in the same class as ∇ , but with torsion \bar{T} , namely

$$\bar{\nabla}_X Y = \nabla_X Y + \frac{1}{2}\bar{T}(X, Y).$$

It is also of interest to inquire when two connections $\bar{\nabla}$ and ∇ have the same geodesics, with possibly *different* parameterizations. Such connections are called “projectively equivalent”, because in ordinary space a projective map is one that takes straight lines to straight lines.

17. PROPOSITION (H. WEYL; 1921). The following are equivalent:

- (a) The connections $\bar{\nabla}$ and ∇ have the same geodesics, with possibly different parameterizations.
- (b) For every X there is λ_X with $D(X, X) = \lambda_X \cdot X$.
- (c) There is a (unique) 1-form ω with $S(X, Y) = \omega(X)Y + \omega(Y)X$.

PROOF. (a) \Rightarrow (b): Given $X_p \in M$, let γ be the geodesic for the connection ∇ with $\gamma' = X_p$, and let $\bar{\gamma}$ be the reparameterization which makes it a geodesic for $\bar{\nabla}$. Let X be a vector field which equals $d\gamma/dt$ along γ and \bar{X} a vector field which equals $d\bar{\gamma}/dt$ along γ . Then along γ we have $X = f\bar{X}$ for some f . So

$$\begin{aligned} D(X_p, X_p) &= \bar{\nabla}_{X_p} X - \nabla_{X_p} X \\ &= \bar{\nabla}_{f(p)\bar{X}_p} f\bar{X} - 0 \\ &= f(p)[\bar{X}_p(f) \cdot \bar{X}_p + f(p)\bar{\nabla}_{\bar{X}_p} \bar{X}] \\ &= f(p)\bar{X}_p(f) \cdot \bar{X}_p = \bar{X}_p(f)X_p. \end{aligned}$$

(b) \Rightarrow (a): Let γ be a geodesic for ∇ , and let X be a vector field which equals $d\gamma/dt$ along γ . Then for $p = \gamma(t)$ we have

$$(1) \quad \bar{\nabla}_{X_p} X = D(X_p, X_p) + \nabla_{X_p} X = \lambda_{X_p} X_p = g(t)X_p, \quad \text{say.}$$

Let $f(t) = e^{G(t)} \neq 0$, where $G'(t) = g(t)$ so that

$$(2) \quad \frac{df}{dt} = g(t)f(t),$$

and let \bar{X} be a vector field such that

$$(3) \quad \bar{X}_{\gamma(t)} = \frac{1}{f(t)} X_{\gamma(t)} = \frac{1}{f(t)} \frac{d\gamma}{dt}.$$

Then

$$\begin{aligned} \bar{\nabla}_{\bar{X}_p} \bar{X} &= \frac{1}{f(t)} \bar{\nabla}_{X_p} \frac{1}{f} X \\ &= \frac{1}{f(t)} \left[X_p \left(\frac{1}{f} \right) X_p + \frac{1}{f(t)} \bar{\nabla}_{X_p} X \right] \\ &= \frac{1}{f(t)} \left[-\frac{df/dt}{f(t)^2} X_p + \frac{1}{f(t)} g(t) X_p \right] && \text{by (1)} \\ &= \frac{1}{f(t)} \left[-\frac{df/dt}{f(t)^2} X_p + \frac{df/dt}{f(t)^2} X_p \right] && \text{by (2)} \\ &= 0. \end{aligned}$$

Consequently, γ is a geodesic for \bar{V} if we reparameterize it as $\bar{\gamma} = \gamma \circ p^{-1}$, where p is chosen so that

$$\frac{d\gamma}{dt} = \bar{X}_{\bar{\gamma}(t)},$$

i.e., so that

$$(p^{-1})'(t) \frac{d\gamma}{dt} \Big|_{p^{-1}(t)} = \frac{1}{f(p^{-1}(t))} \frac{d\gamma}{dt} \Big|_{p^{-1}(t)} \quad \text{by (3).}$$

For this we need

$$\frac{1}{p'(p^{-1}(t))} = \frac{1}{f(p^{-1}(t))},$$

or simply $p'(s) = f(s) > 0$ (compare Problem I.9-27).

(c) \Rightarrow (b) is clear.

(b) \Rightarrow (c): We first establish the following algebraic

LEMMA. Let V be a vector space, and $S: V \times V \rightarrow V$ a symmetric bilinear map such that for each $v \in V$ there is $\lambda_v \in \mathbb{R}$ with

$$S(v, v) = \lambda_v v.$$

Then there is a unique $\phi \in V^*$ such that

$$S(v, w) = \phi(v)w + \phi(w)v.$$

PROOF. If ϕ exists, clearly $\phi(v) = \lambda_v/2$ for $v \neq 0$. Conversely, if this definition makes ϕ linear we will be done, for then

$$\begin{aligned} S(v, v) + S(w, w) + 2S(v, w) &= S(v + w, v + w) = \lambda_{v+w}(v + w) \\ &= (\lambda_v + \lambda_w)(v + w), \end{aligned}$$

so

$$2\phi(v)v + 2\phi(w)w + 2S(v, w) = 2\phi(v)v + 2\phi(w)w + 2\phi(v)w + 2\phi(w)v,$$

which yields

$$S(v, w) = \phi(v)w + \phi(w)v.$$

We prove that ϕ is linear as follows. From

$$\begin{aligned} \lambda_{v+w}(v + w) &= \lambda_v v + \lambda_w w + 2S(v, w) \\ \lambda_{v-w}(v - w) &= \lambda_v v + \lambda_w w - 2S(v, w) \end{aligned}$$

we obtain

$$(\lambda_{v+w} + \lambda_{v-w} - 2\lambda_v)v + (\lambda_{v+w} - \lambda_{v-w} - 2\lambda_w)w = 0.$$

For linearly independent v and w we thus have

$$\begin{aligned}\lambda_{v+w} + \lambda_{v-w} - 2\lambda_v &= 0 \\ \lambda_{v+w} - \lambda_{v-w} - 2\lambda_w &= 0,\end{aligned}$$

and hence $\lambda_{v+w} = \lambda_v + \lambda_w$. This is clearly also true if v and w are linearly dependent. Homogeneity is likewise trivial. **Q.E.D.**

When V is n -dimensional, we have the explicit formula

$$(*) \quad \phi(v) = \frac{1}{n+1} \text{ trace of } w \mapsto S(v, w),$$

which can be deduced as follows. Choose a basis v_1, \dots, v_n with $v = v_1$. Then

$$S(v_1, v_j) = \phi(v_1)v_j + \phi(v_j)\phi_1,$$

so

$$\text{trace of } w \mapsto S(v_1, w) = \sum_{j=1}^n j^{\text{th}} \text{ component of } \phi(v_1)v_j + \phi(v_j)v_1;$$

this component is $\phi(v_1)$ for $j \neq 1$, and $2\phi(v_1)$ for $j = 1$.

Formula $(*)$ clearly shows that

$$\omega(X) = \frac{1}{n+1} \text{ trace } Y \mapsto S(X, Y)$$

is the desired C^∞ 1-form on M . \blacklozenge

18. COROLLARY. The connections $\bar{\nabla}$ and ∇ have the same torsion and the same geodesics (suitably reparameterized) if and only if there is a (unique) 1-form ω with

$$D(X, Y) = \omega(X)Y + \omega(Y)X.$$

Note, by the way, that for any connection ∇ and any 1-form ω , the function

$$\bar{\nabla}_X Y = \nabla_X Y + \omega(X)Y + \omega(Y)X$$

always is a connection, with the same torsion as ∇ .

ADDENDUM 2

RIEMANN'S INVARIANT DEFINITION
OF THE CURVATURE TENSOR

In part C of Chapter 4 we presented an extract from Riemann's paper of 1861. Our use of "..." near the end was rather deceitful, because one of the omissions was by no means a minor one. In the deleted portion Riemann gives a result, which, although it plays no further role in our development of Riemannian geometry, is nevertheless extremely interesting, for it amounts to another invariant definition of the curvature tensor. We therefore give below an unabridged version of the second part of the extract, beginning with the last paragraph in the version in Chapter 4C. It will present greater difficulties of interpretation than any thing else we have read, and is followed immediately by an exposition in modern terms.

* * * * *

The functions g_{ij} must necessarily satisfy these equations whenever $\sum_{i,j} g_{ij} dy^i dy^j$ can be transformed into the form $\sum_i (dx^i)^2$: we denote the left side of this equation by

$$(ij,kl).$$

To make the nature of this equation more transparent, we form the expression

$$\delta\delta \sum g_{ij} dy^i dy^j - 2d\delta \sum g_{ij} dy^i \delta y^j + dd \sum g_{ij} \delta y^i \delta y^j,$$

the variations of the second order d^2 , $d\delta$, δ^2 being so determined that

$$\delta' \sum g_{ij} dy^i \delta y^j - \delta \sum g_{ij} dy^i \delta' y^j - d \sum g_{ij} \delta y^i \delta' y^j = 0$$

$$\delta' \sum g_{ij} dy^i dy^j - 2d \sum g_{ij} dy^i \delta' y^j = 0$$

$$\delta' \sum g_{ij} \delta y^i \delta y^j - 2\delta \sum g_{ij} \delta y^i \delta' y^j = 0,$$

δ' denoting an arbitrary variation. In this way the above expression becomes

$$(II) \quad = \sum (ij,kl)(dy^i \delta y^j - dy^j \delta y^i)(dy^k \delta y^l - dy^l \delta y^k).$$

Now from the formation of this expression it is immediately evident that a change of the independent variables changes it into a new form depending in

the same way on the $\sum g_{ij} dx^i dx^j$. And if the quantities g_{ij} are constant, all coefficients of the expression (II) turn out to be equal to 0. Thus if $\sum g_{ij} dy^i dy^j$ can be transformed into a similar expression with constant coefficients, it is necessary that expression (II) vanishes identically.

In the same way it turns out that if the expression (II) does not vanish, the expression

$$(III) \quad -\frac{1}{2} \frac{\sum (ij,kl)(dy^i \delta y^j - dy^j \delta y^i)(dy^k \delta y^l - dy^l \delta y^k)}{\sum g_{ij} dy^i dy^j \sum g_{ij} dy^i \delta y^j - \left(\sum g_{ij} dy^i \delta y^j\right)^2}$$

does not change if the independent variables are changed, and moreover remains unchanged if in place of the variations dy^i , δy^i , arbitrary independent linear expressions of them, $\alpha dy^i + \beta \delta y^i$, $\gamma dy^i + \delta y^i$ are substituted. Moreover, the maximum and minimum values of the function (III) of the same dy^i , δy^i depend neither on the form of the expression $\sum g_{ij} dy^i dy^j$ nor on the values of the variations dy^i , δy^i , whence from these values it can be determined when two expressions of this kind can be transformed into each other.

These interpretations can be illustrated by what one might call a geometrical interpretation, which, although it depends on unusual conceptions, it will nevertheless help, as one goes along, to have touched upon.

The expression $\sqrt{\sum g_{ij} dy^i dy^j}$ can be regarded as just the line element in a generalized space of n dimensions transcending our intuition. If in this space at the point (y^1, \dots, y^n) all shortest lines are drawn, in which the initial elements of variation of the y^i are $\alpha dy^1 + \beta \delta y^1 : \alpha dy^2 + \beta \delta y^2 : \dots : \alpha dy^n + \beta \delta y^n$, α and β denoting arbitrary quantities, these lines make up a surface which can be developed in the space of our common intuition. In this way the expression (III) will measure the curvature of this surface at the point (y^1, \dots, y^n) .

If we now return to the case $n = 3$, the expression (II) is a form of the second degree in

$$dy^2 \delta y^3 - dy^3 \delta y^2, \quad dy^3 \delta y^1 - dy^1 \delta y^3, \quad dy^1 \delta y^2 - dy^2 \delta y^1,$$

and in this case we obtain six equations, which the functions g_{ij} are required to satisfy if $\sum g_{ij} dy^i dy^j$ can be transformed into a form with constant coefficients. Given an acquaintance with the traditional methods, it is demonstrated without difficulty that these six conditions, when they are satisfied, suffice. It is to be observed nevertheless that only three of them are independent.

Riemann's description of the curvature tensor is particularly difficult to decipher because he is using classical notation from the calculus of variations. To state it in modern terms, consider a function $s: \mathbb{R}^2 \rightarrow M$, with $s(0) = p \in M$. This function s can be thought of as a "2-parameter variation of the point p ", and on M it gives rise to 2 vector fields,

$$\begin{aligned}\frac{\partial s}{\partial x} &= s_* \left(\frac{\partial}{\partial x} \right) \\ \frac{\partial s}{\partial y} &= s_* \left(\frac{\partial}{\partial y} \right).\end{aligned}$$

This notation involves the usual ambiguities, and in conformity with this, the function

$$(x, y) \mapsto \left\langle \frac{\partial s}{\partial x}(x, y), \frac{\partial s}{\partial y}(x, y) \right\rangle$$

on \mathbb{R}^2 will be denoted simply by

$$\left\langle \frac{\partial s}{\partial x}, \frac{\partial s}{\partial y} \right\rangle.$$

Riemann directs us to consider the expression

$$(A) \quad \frac{\partial^2}{\partial x^2} \left\langle \frac{\partial s}{\partial y}, \frac{\partial s}{\partial y} \right\rangle - 2 \frac{\partial^2}{\partial x \partial y} \left\langle \frac{\partial s}{\partial x}, \frac{\partial s}{\partial y} \right\rangle + \frac{\partial^2}{\partial y^2} \left\langle \frac{\partial s}{\partial x}, \frac{\partial s}{\partial x} \right\rangle.$$

If we let

$$X = \frac{\partial s}{\partial x}, \quad Y = \frac{\partial s}{\partial y},$$

then the value of this expression at $(x, y) = (0, 0)$ can be written

$$X_p(X(\langle Y, Y \rangle)) - 2X_p(Y(\langle X, Y \rangle)) + Y_p(Y(\langle X, X \rangle)),$$

or simply

$$(B) \quad [XX\langle Y, Y \rangle - 2XY\langle X, Y \rangle + YY\langle X, X \rangle](p).$$

On the other hand, to be certain that the expression (B) will equal (A) for some $s: \mathbb{R}^2 \rightarrow M$, we need to know that $[X, Y] = 0$. Riemann does not consider all $s: \mathbb{R}^2 \rightarrow M$, but only those with the following property: If $\sigma: \mathbb{R}^3 \rightarrow M$ is $\sigma(x, y, 0) = s(x, y)$, then

$$\left. \begin{aligned} \frac{\partial}{\partial z} \left\langle \frac{\partial \sigma}{\partial x}, \frac{\partial \sigma}{\partial y} \right\rangle - \left\langle \frac{\partial}{\partial y} \frac{\partial \sigma}{\partial x}, \frac{\partial \sigma}{\partial z} \right\rangle - \left\langle \frac{\partial}{\partial x} \frac{\partial \sigma}{\partial y}, \frac{\partial \sigma}{\partial z} \right\rangle &= 0 \\ \frac{\partial}{\partial z} \left\langle \frac{\partial \sigma}{\partial x}, \frac{\partial \sigma}{\partial x} \right\rangle - 2 \frac{\partial}{\partial x} \left\langle \frac{\partial \sigma}{\partial x}, \frac{\partial \sigma}{\partial z} \right\rangle &= 0 \\ \frac{\partial}{\partial z} \left\langle \frac{\partial \sigma}{\partial y}, \frac{\partial \sigma}{\partial y} \right\rangle - 2 \frac{\partial}{\partial y} \left\langle \frac{\partial \sigma}{\partial y}, \frac{\partial \sigma}{\partial z} \right\rangle &= 0 \end{aligned} \right\} \quad \text{at } (0, 0, 0).$$

Riemann claims that if $s: \mathbb{R}^2 \rightarrow M$ has this property, then the value of (A) at $(x, y) = (0, 0)$ is $-2\langle R(X_p, Y_p)Y_p, X_p \rangle$. (To account for the factor of -2 recall that Riemann's (ij, kl) equals $R_{ijkl}/2 = -R_{ijkl}/2$, and note that Riemann has $(dy^i \wedge dy^j) \otimes (dy^k \wedge dy^l)$ instead of $dy^i \otimes dy^j \otimes dy^k \otimes dy^l$.) This assertion can be rephrased as follows.

19. THEOREM (RIEMANN). For any linearly independent vectors $X_p, Y_p \in M_p$ there are vector fields X, Y extending them such that

(a) $[X, Y] = 0$

(b) For every vector field Z with $[X, Z](p) = [Y, Z](p) = 0$ we have

$$\left. \begin{array}{l} (1) \quad Z\langle X, Y \rangle - Y\langle X, Z \rangle - X\langle Y, Z \rangle = 0 \\ (2) \quad Z\langle X, X \rangle - 2X\langle X, Z \rangle = 0 \\ (3) \quad Z\langle Y, Y \rangle - 2Y\langle Y, Z \rangle = 0 \end{array} \right\} \quad \text{at } p.$$

For any such vector fields X and Y we have

$$-2\langle R(X_p, Y_p)Y_p, X_p \rangle = [XX\langle Y, Y \rangle - 2XY\langle X, Y \rangle + YY\langle X, X \rangle](p).$$

PROOF. Using Corollary 7, we see that equation (2) is equivalent to

$$2\langle \nabla_Z X, X \rangle - 2\langle \nabla_X X, Z \rangle - 2\langle X, \nabla_X Z \rangle = 0 \quad \text{at } p$$

and hence to

$$\langle \nabla_X X, Z \rangle = 0 \quad \text{at } p,$$

since $\nabla_Z X(p) - \nabla_X Z(p) = [X, Z](p) = 0$. Thus equation (2) is equivalent to

$$(2') \quad \nabla_X X(p) = 0.$$

Similarly, (3) is equivalent to

$$(3') \quad \nabla_Y Y(p) = 0.$$

Finally, (1) is equivalent to the equality, at p ,

$$\langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle - \langle \nabla_Y X, Z \rangle - \langle X, \nabla_Y Z \rangle - \langle \nabla_X Y, Z \rangle - \langle Y, \nabla_X Z \rangle = 0,$$

and hence to

$$\langle \nabla_X Y + \nabla_Y X, Z \rangle = 0 \quad \text{at } p.$$

This is equivalent to

$$0 = \nabla_X Y(p) + \nabla_Y X(p) = 2\nabla_X Y(p), \quad \text{since } \nabla_X Y - \nabla_Y X = [X, Y] = 0,$$

so (1) is equivalent to

$$(1') \quad \nabla_X Y(p) = \nabla_Y X(p) = 0.$$

We can obtain such vector fields X and Y , with given values $X_p, Y_p \in M_p$, by mapping \mathbb{R}^2 into M in such a way that the x and y axes go into two geodesics in M with $s_*(\partial/\partial y)$ parallel along the image of the x -axis and $s_*(\partial/\partial x)$ parallel along the image of the y -axis.

Now for such vector fields we have

$$\begin{aligned} XX\langle Y, Y \rangle(p) &= 2X\langle \nabla_X Y, Y \rangle(p) = 2\langle \nabla_X \nabla_X Y, Y \rangle(p) + 2\langle \nabla_X Y, \nabla_X Y \rangle(p) \\ &= 2\langle \nabla_X \nabla_X Y, Y \rangle(p) \\ YY\langle X, X \rangle(p) &= 2\langle \nabla_Y \nabla_Y X, X \rangle(p) \\ -2XY\langle X, Y \rangle(p) &= -2X(\langle \nabla_Y X, Y \rangle + \langle X, \nabla_Y Y \rangle)(p) \\ &= -2\langle \nabla_X \nabla_Y X, Y \rangle(p) - 2\langle \nabla_Y X, \nabla_X Y \rangle(p) \\ &\quad - 2\langle \nabla_X X, \nabla_Y Y \rangle(p) - 2\langle X, \nabla_X \nabla_Y Y \rangle(p) \\ &= -2\langle \nabla_X \nabla_Y X, Y \rangle(p) - 2\langle X, \nabla_X \nabla_Y Y \rangle(p). \end{aligned}$$

So the sum is

$$\begin{aligned} &2[\langle \nabla_X \nabla_X Y, Y \rangle - \langle \nabla_X \nabla_Y X, Y \rangle](p) + 2[\langle \nabla_Y \nabla_Y X, X \rangle - \langle \nabla_X \nabla_Y Y, X \rangle](p) \\ &= 0 + 2[\langle \nabla_Y \nabla_X Y, X \rangle - \langle \nabla_X \nabla_Y Y, X \rangle](p) \\ &\quad \text{since } \nabla_Y X - \nabla_X Y = [X, Y] = 0 \text{ everywhere} \\ &= 2\langle R(Y_p, X_p)Y_p, X_p \rangle \\ &= -2\langle R(X_p, Y_p)Y_p, X_p \rangle. \quad \spadesuit \end{aligned}$$

CHAPTER 7

THE REPÈRE MOBILE (THE MOVING FRAME)

The previous chapter betrayed the historical development which we have been following, for the ∇ operator did not appear until very late in the game, around 1954.* In the meantime, Élie Cartan had elaborated a completely different theory, the method of the repère mobile. Despite the fact that this theory was invented soon after the Ricci calculus, some of its features are most easily understood by referring to the ∇ operator which historically came so much later.

Roughly speaking, the relationship between the results of Chapter 5 and those of Chapter 6 can be characterized as follows. When working with ∇ operators we express results in terms of arbitrary vector fields, while in the classical theory we always use the vector fields $X_i = \partial/\partial x^i$ given by a coordinate system (x, U) . At each point $p \in U$, these vector fields provide us with an ordered basis

$$(X_1(p), \dots, X_n(p)) \quad \text{for } M_p.$$

In general, an ordered basis (v_1, \dots, v_n) for a vector space V will also be called a **frame** in V . Now the vector fields X_i may be used to determine a function

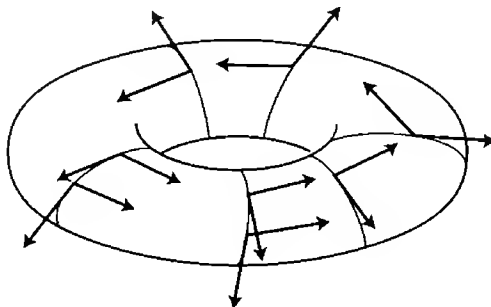
$$p \mapsto (X_1(p), \dots, X_n(p)),$$

whose values are frames in the various tangent spaces M_p ; such a function is called a **moving frame**. Conversely, every moving frame $p \mapsto F_p$ determines vector fields X_1, \dots, X_n by $F_p = (X_1(p), \dots, X_n(p))$. Consequently, as a matter of convenience we will not distinguish between everywhere linearly independent vector fields X_1, \dots, X_n , and the moving frame they determine. In É. Cartan's theory, the basic idea, whose ramifications turn out to be extremely significant, is to express everything in terms of an arbitrary moving frame X_1, \dots, X_n , and not just in terms of the "natural moving frame" $X_i = \partial/\partial x^i$ given by a coordinate system x .

Notice that a moving frame X_1, \dots, X_n need not be the natural frame for any coordinate system, since we need not have $[X_i, X_j] = 0$. Also, a moving frame

*Nomizu, *Invariant affine connections on homogeneous spaces*, Amer. J. Math. **76** (1954), 33–65.

may exist on a region which cannot even be included in a coordinate system. For example, there is a moving frame X_1, X_2 on the whole torus. On a Riemannian



manifold $(M, \langle \cdot, \cdot \rangle)$ we define an **orthonormal** moving frame X_1, \dots, X_n to be one such that each $(X_1(p), \dots, X_n(p))$ is an orthonormal frame for M_p . The frame illustrated above, on the torus, is an example. An arbitrary moving frame X_1, \dots, X_n gives rise to an orthonormal moving frame if we apply the Gram-Schmidt orthonormalization process to it (Problem I.9-11).

Given a moving frame X_1, \dots, X_n on (an open subset U of) M , we now ask if it is possible to describe quantitatively just how the frame is moving. As a guide, we first consider a moving frame X_1, \dots, X_n in \mathbb{R}^n . In this case, we can, with a little abuse of notation, consider each X_i as an \mathbb{R}^n -valued function $X_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$. If we consider the identity map as an \mathbb{R}^n -valued function $P: \mathbb{R}^n \rightarrow \mathbb{R}^n$, then the \mathbb{R}^n -valued 1-form dP is just

$$dP(X_a) = X.$$

Consequently, if we introduce 1-forms θ^i by

$$(I) \quad dP = \sum_{i=1}^n \theta^i \cdot X_i \quad \text{i.e.,} \quad dP(X_a) = \sum_{i=1}^n \theta^i(X_a) \cdot X_i(a) \in \mathbb{R}^n,$$

then θ^i are just the “dual forms” to the X_i , that is, $\theta^i(X_j) = \delta_j^i$. We also introduce 1-forms ω_j^i by

$$(II) \quad dX_j = \sum_{i=1}^n \omega_j^i \cdot X_i.$$

Clearly $\omega_j^i(X_a)$ is the X_i component of $dX_j(X_a)$. Since $dX_j(X_a)$ is just the directional derivative of X_j in the direction X_a , we can interpret $\omega_j^i(X_a)$ as the *rate at which X_j rotates toward $X_i(a)$ as we move along a curve with tangent vector X_a .*

Now it is not possible to find a moving frame with arbitrary θ^i and ω_j^i ; certain integrability conditions must be satisfied:

1. PROPOSITION. The 1-forms θ^i and ω_j^i for a moving frame X_1, \dots, X_n in \mathbb{R}^n satisfy the **structural equations of Euclidean space**:

$$\begin{aligned} d\theta^i &= \sum_k \theta^k \wedge \omega_k^i = - \sum_k \omega_k^i \wedge \theta^k \\ d\omega_j^i &= - \sum_k \omega_k^i \wedge \omega_j^k. \end{aligned}$$

PROOF. Noting that the equation $dP = \sum \theta^i \cdot X_i$ can also be written $dP = \sum_i \theta^i \wedge X_i$ (for the \mathbb{R}^n -valued 0-form X_i), we have

$$\begin{aligned} 0 = d^2P &= d\left(\sum_i \theta^i \wedge X_i\right) = \sum_i d\theta^i \wedge X_i - \sum_k \theta^k \wedge dX_k \\ &= \sum_i d\theta^i X_i - \sum_k \theta^k \wedge \sum_i \omega_k^i X_i \quad \text{by (II).} \end{aligned}$$

Setting the coefficient of each X_i equal to 0, we obtain the first structural equation.

We also have

$$\begin{aligned} 0 = d^2X_j &= \sum_i d\omega_j^i X_i - \sum_k \omega_j^k \wedge dX_k \\ &= \sum_i d\omega_j^i X_i - \sum_k \omega_j^k \wedge \sum_i \omega_k^i X_i, \end{aligned}$$

from which we immediately deduce the second structural equation. ♦

The structural equations can be written much more compactly if we use slightly modified matrix notation. Henceforth, we will write matrices as $A = (A_j^i)$, and define

$$(A \cdot B)_j^i = \sum_k A_k^i B_j^k,$$

so that our new A_k^i corresponds to the old A_{ik} . If $\mathbf{v} = (v_1, \dots, v_n)$ is an ordered n -tuple of vectors, and we define

$$w_j = \sum_i A_j^i v_i,$$

then the n -tuple $\mathbf{w} = (w_1, \dots, w_n)$ will be denoted by $\mathbf{v} \cdot A$. We put the A on the right because we have

$$\begin{aligned} [\mathbf{v} \cdot (A \cdot B)]_j &= \sum_i (A \cdot B)_j^i v_i = \sum_i \sum_k A_k^i B_j^k v_i \\ &= \sum_k B_j^k \left(\sum_i A_k^i v_i \right) = \sum_k B_j^k (\mathbf{v} \cdot A)_k \\ &= [(\mathbf{v} \cdot A) \cdot B]_j, \end{aligned}$$

which we can write simply as*

$$\mathbf{v} \cdot (AB) = (\mathbf{v} \cdot A) \cdot B.$$

Naturally, if $\mathbf{X} = (X_1, \dots, X_n)$ is an n -tuple of vector fields on a manifold M , and $A = (A_j^i)$ is a matrix of functions (or a matrix-valued function, whichever way you prefer to look at it), then $\mathbf{X} \cdot A$ denotes the n -tuple of vector fields $(\mathbf{X} \cdot A)_j = \sum_i A_j^i X_i$.

We extend this notation to forms in the natural way. If $\omega = (\omega_j^i)$ and $\eta = (\eta_j^i)$ are matrices of forms, of degree k and l , respectively, then $\omega \wedge \eta$ is the matrix of $(k + l)$ -forms

$$(\omega \wedge \eta)_j^i = \sum_k \omega_k^i \wedge \eta_j^k.$$

If θ denotes the column vector of l -forms $\theta^1, \dots, \theta^n$, then $\omega \wedge \theta$ is the column vector of $(k + l)$ -forms

$$(\omega \wedge \theta)^i = \sum_j \omega_j^i \wedge \theta^j.$$

With these conventions we can write the structural equations of Euclidean space as

$$\begin{aligned} d\theta &= -\omega \wedge \theta \\ d\omega &= -\omega \wedge \omega. \end{aligned}$$

Henceforth, we will use this notation whenever convenient; for a while the reader may feel more secure rewriting things in standard form. Our very next result justifies our characterization of the structural equations as “integrability conditions”.

*We express this by saying that the $n \times n$ matrices **act on the right** on the set of all n -tuples of vectors of V . On the other hand, suppose we choose a *fixed* basis v_1, \dots, v_n for V , and then define $A \cdot v$ for $v \in V$ by defining $A \cdot v_j = \sum_i A_j^i v_i$ and extending by linearity. In this case we will have $(A \cdot B) \cdot v = A \cdot (B \cdot v)$; so we say that the $n \times n$ matrices **act on the left** on V .

2. PROPOSITION. Let $\omega = (\omega_j^i)$ be a matrix of 1-forms on \mathbb{R}^n which satisfy the second structural equations

$$d\omega = -\omega \wedge \omega \quad \text{i.e.,} \quad d\omega_j^i = -\sum_k \omega_k^i \wedge \omega_j^k.$$

Then,

(1) In a neighborhood of 0 there is a matrix $A = (A_j^i)$ of functions, with arbitrary initial condition $A(0)$, such that

$$dA = -\omega \wedge A \quad \text{i.e.,} \quad dA_j^i = -\sum_k \omega_k^i A_j^k.$$

(2) In a neighborhood of 0 there is a moving frame X_1, \dots, X_n with arbitrary initial conditions $X_1(0), \dots, X_n(0)$, so that the dual 1-forms θ^i satisfy the first structural equation

$$d\theta = -\omega \wedge \theta \quad \text{i.e.,} \quad d\theta^i = -\sum_k \omega_k^i \wedge \theta^k.$$

PROOF. (1) Let y^1, \dots, y^n be the standard coordinate system on \mathbb{R}^n , and let y^1, \dots, y^n, z_j^i be the standard coordinate system on \mathbb{R}^{n+n^2} . Let Z be the matrix of functions $Z = (z_j^i)$ and consider the matrix of 1-forms

$$(*) \quad \Lambda = dZ + (\omega \wedge Z).$$

We have

$$\begin{aligned} d\Lambda &= d\omega \wedge Z - \omega \wedge dZ \\ &= -(\omega \wedge \omega) \wedge Z - \omega \wedge [\Lambda - (\omega \wedge Z)] \quad \text{by } (*) \text{ and the hypothesis} \\ &= -\omega \wedge \Lambda. \end{aligned}$$

By Proposition I.7-14, the n -dimensional distribution

$$\Delta_p = \bigcap_{i,j=1}^n \ker \Lambda_j^i(p)$$

in \mathbb{R}^{n+n^2} is integrable. Since $\Delta_{(0,z_0)} = (y^1, \dots, y^n)$ -plane, the integral manifold though any point $(0, z_0)$ is locally the graph of a function $p \mapsto A(p) \in \mathbb{R}^{n^2}$ with

$A(0) = z_0$. Since $dZ + (\omega \wedge Z) = 0$ on this graph, we conclude, as in the proof of Theorem I.10-17, that

$$(**) \quad dA = -\omega \wedge A.$$

(2) Choose A to be non-singular at 0, and define 1-forms θ^i by

$$\theta^i = \sum_j A_j^i dy^j,$$

which we can write simply as

$$\theta = A \wedge dy.$$

Then

$$\begin{aligned} d\theta &= dA \wedge dy \\ &= -\omega \wedge A \wedge dy \quad \text{by } (**) \\ &= -\omega \wedge \theta. \end{aligned}$$

So we just define the moving frame X_1, \dots, X_n by $\theta^i(X_j) = \delta_j^i$. ♦

For orthonormal moving frames we have one more relation:

3. PROPOSITION. The forms ω_j^i for an orthonormal moving frame X_1, \dots, X_n in \mathbb{R}^n satisfy

$$\omega_j^i = -\omega_i^j,$$

i.e., the matrix ω is skew-symmetric.

PROOF. We have

$$0 = d(\langle X_i, X_j \rangle) = \langle dX_i, X_j \rangle + \langle X_i, dX_j \rangle.$$

Here $\langle dX_i, X_j \rangle(X_a)$ means $\langle dX_i(X_a), X_j(a) \rangle$. Since the X_i are orthogonal, clearly $\langle dX_i, X_j \rangle(X_a)$ is just the X_j component of $dX_i(X_a)$. This means that $\langle dX_i, X_j \rangle = \omega_j^i$. ♦

Now consider a moving frame X_1, \dots, X_n on a manifold M . We can still define the **dual 1-forms** θ^i by $\theta^i(X_j) = \delta_j^i$; equation (I) can now be rewritten as $X_q = \sum_i \theta^i(X_q) \cdot X_i(q)$ for any $X_q \in M_q$, or simply $dP = \sum_i \theta^i \cdot X_i$, where “ dP ” denotes the identity map of a tangent space into itself. We cannot use equation (II) to define the forms ω_j^i , since “ dX_i ” makes no sense on a general manifold. However, Propositions 1 and 3 suggest a way of defining these forms on a Riemannian manifold.

4. PROPOSITION. Let X_1, \dots, X_n be a moving frame on a manifold M , and let θ^i be the dual 1-forms. Then there exist unique 1-forms ω_j^i such that

$$\begin{aligned} \text{(a)} \quad & \omega_j^i = -\omega_i^j \\ \text{(b)} \quad & d\theta^i = \sum_k \theta^k \wedge \omega_k^i. \end{aligned}$$

PROOF. Suppose ω_j^i satisfy (a) and (b). There are unique functions a_{jk}^i and b_{jk}^i with

$$\begin{aligned} \omega_j^i &= \sum_k a_{jk}^i \theta^k, \\ d\theta^i &= \frac{1}{2} \sum_{j,k} b_{jk}^i \theta^j \wedge \theta^k, \quad b_{jk}^i = -b_{kj}^i. \end{aligned}$$

Then (a) is equivalent to

$$\text{(a')} \quad a_{ik}^j = -a_{jk}^i,$$

while (b) gives

$$\frac{1}{2} \sum_{j,k} b_{jk}^i \theta^j \wedge \theta^k = d\theta^i = \sum_j \theta^j \wedge \omega_j^i = \sum_{j,k} a_{jk}^i \theta^j \wedge \theta^k,$$

and hence

$$\text{(b')} \quad a_{jk}^i - a_{kj}^i = b_{jk}^i.$$

Cyclically permuting i, j, k , we obtain from (a') and (b')

$$a_{jk}^i = \frac{1}{2}(b_{jk}^i + b_{ki}^j - b_{ij}^k).$$

This proves uniqueness.

Now suppose we define ω_j^i by

$$\omega_j^i = \sum_k a_{jk}^i \theta^k,$$

where the a_{jk}^i are as defined above, and hence satisfy (b'). It is easy to check that equation (a') holds, and hence that equation (a) holds. Moreover, from equation (b') we have

$$\omega_j^i(X_k) - \omega_k^i(X_j) = a_{jk}^i - a_{kj}^i = b_{jk}^i;$$

it follows that

$$\begin{aligned}
 \sum_l \theta^l \wedge \omega_l^i(X_j, X_k) &= \sum_l \theta^l(X_j) \omega_l^i(X_k) - \theta^l(X_k) \omega_l^i(X_j) \\
 &= \sum_l \delta_j^l \omega_l^i(X_k) - \delta_k^l \omega_l^i(X_j) \\
 &= \omega_j^i(X_k) - \omega_k^i(X_j) \\
 &= b_{jk}^i,
 \end{aligned}$$

which is equivalent to equation (b). ♦

Although we have used Proposition 3 to motivate Proposition 4, the latter result seems to involve neither a Riemannian metric nor an orthonormal moving frame. However the two are, in a sense, really there, since there is a unique Riemannian metric on the domain of the moving frame X_1, \dots, X_n which makes it an orthonormal moving frame. Naturally, if we are already given a Riemannian metric $\langle \cdot, \cdot \rangle$ on M , then we will expect the 1-forms ω_j^i to have some significance for this metric only when the moving frame is orthonormal with respect to it.

Let us therefore consider an *orthonormal* moving frame X_1, \dots, X_n on a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$. The unique 1-forms ω_j^i given by Proposition 4 are called the **connection forms** for the moving frame X_1, \dots, X_n . They satisfy $\omega_j^i = -\omega_i^j$ and the first structural equation, by their very definition. On the other hand, there is no reason to expect them to satisfy the second structural equation. Recognizing this, we define a matrix of 2-forms $\Omega = (\Omega_j^i)$ by

$$d\omega = -\omega \wedge \omega + \Omega \quad \text{i.e.,} \quad d\omega_j^i = - \sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i.$$

The 2-forms Ω_j^i are called the **curvature forms** for the orthonormal moving frame X_1, \dots, X_n . The names “connection forms” and “curvature forms” are explained by the very next theorem. Let us set

$$\begin{aligned}
 \nabla_{X_i} X_j &= \sum_{k=1}^n \Gamma_{ij}^k X_k \\
 R(X_i, X_j) X_k &= \sum_{l=1}^n \mathbf{R}_{kij}^l X_l,
 \end{aligned}$$

where ∇ is the Levi-Civita connection for $\langle \cdot, \cdot \rangle$, and R is its curvature tensor. We use bold letters $\mathbf{\Gamma}$ and \mathbf{R} to remind ourselves that these are not components with respect to a coordinate system; thus, for example, we do *not* necessarily have $\mathbf{\Gamma}_{ij}^k = \mathbf{\Gamma}_{ji}^k$ (but we clearly do have $\mathbf{R}_{kij}^l = \mathbf{R}_{kji}^l$).

5. THEOREM. Let X_1, \dots, X_n be an *orthonormal* moving frame on a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, and let $\theta^i, \omega_j^i, \Omega_j^i$ be the dual forms, connection forms, and curvature forms for this moving frame. Then we have the **structural equations of $(M, \langle \cdot, \cdot \rangle)$** :

$$\begin{aligned} d\theta^i &= - \sum_k \omega_k^i \wedge \theta^k & d\theta &= -\omega \wedge \theta \\ d\omega_j^i &= - \sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i & d\omega &= -\omega \wedge \omega + \Omega, \end{aligned}$$

where

$$\omega_j^i = \sum_k \Gamma_{kj}^i \theta^k, \quad \Omega_j^i = \frac{1}{2} \sum_{k,l} \mathbf{R}_{jkl}^i \theta^k \wedge \theta^l = \sum_{k < l} \mathbf{R}_{jkl}^i \theta^k \wedge \theta^l.$$

PROOF. By the uniqueness part of Proposition 4, we can prove the first structural equation by showing that if we define $\omega_j^i = \sum_k \Gamma_{kj}^i \theta^k$, then the ω_j^i do satisfy conditions (a) and (b) of that proposition. Now by Corollary 6-7 we have

$$\begin{aligned} 0 &= X_k \langle X_i, X_j \rangle = \langle \nabla_{X_k} X_i, X_j \rangle + \langle X_i, \nabla_{X_k} X_j \rangle \\ &= \Gamma_{ki}^j + \Gamma_{kj}^i; \end{aligned}$$

this immediately implies that $\omega_j^i = -\omega_i^j$, which is condition (a).

As in the proof of Proposition 4, we have

$$\sum_l \theta^l \wedge \omega_l^i(X_j, X_k) = \omega_j^i(X_k) - \omega_k^i(X_j),$$

while we also have

$$\begin{aligned} d\theta^i(X_j, X_k) &= X_j(\theta^i(X_k)) - X_k(\theta^i(X_j)) - \theta^i([X_j, X_k]) \\ &\quad \text{by Theorem I.7-13} \\ &= 0 - 0 - \theta^i(\nabla_{X_j} X_k - \nabla_{X_k} X_j) \\ &= \Gamma_{jk}^i - \Gamma_{kj}^i \\ &= \omega_j^i(X_k) - \omega_k^i(X_j); \end{aligned}$$

this proves condition (b).

For the second structural equation we expand

$$\sum_i \mathbf{R}_{jkl}^i X_i = R(X_k, X_l)X_j = \nabla_{X_k} \nabla_{X_l} X_j - \nabla_{X_l} \nabla_{X_k} X_j - \nabla_{[X_k, X_l]} X_j$$

to obtain

$$\mathbf{R}^i_{jkl} = \sum_{\mu} (\Gamma^i_{k\mu} \Gamma^{\mu}_{lj} - \Gamma^i_{l\mu} \Gamma^{\mu}_{kj}) + X_k(\Gamma^i_{lj}) - X_l(\Gamma^i_{kj}) - \theta^i(\nabla_{[X_k, X_l]} X_j).$$

Comparing with

$$\begin{aligned} \left[d\omega_j^i + \sum_{\mu} \omega_{\mu}^i \wedge \omega_j^{\mu} \right] (X_k, X_l) &= X_k(\omega_j^i(X_l)) - X_l(\omega_j^i(X_k)) - \omega_j^i([X_k, X_l]) \\ &\quad + \sum_{\mu} [\omega_{\mu}^i(X_k) \omega_j^{\mu}(X_l) - \omega_{\mu}^i(X_l) \omega_j^{\mu}(X_k)], \end{aligned}$$

we see that this does indeed equal $\mathbf{R}^i_{jkl} = \Omega_j^i(X_k, X_l)$. ♦

Notice that the results of Theorem 5 can also be written

$$\begin{aligned} \nabla_{X_k} X_j &= \sum_i \omega_j^i(X_k) X_i \\ R(X_k, X_l) X_j &= \sum_i \Omega_j^i(X_k, X_l) X_i. \end{aligned}$$

If we did not already have the ∇ operator and its curvature tensor, then we could, and É. Cartan did, use these equations to *define* R . Since the θ^i , hence the ω_j^i , and finally the Ω_j^i , all depend on the moving frame, it is then necessary to check that the resulting definition of R depends only on the values of X_j, X_k, X_l at a given point. We save until the end of the chapter some remarks about what is involved in that. At the moment, we would like to point out that we are already in a position to prove the test case, and thus begin to justify the epithet “structural equations”.

6. THEOREM (THE TEST CASE; FOURTH VERSION). Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. Let Y_1, \dots, Y_n be an orthonormal moving frame around $p \in M$, and let θ^i and ω_j^i be its dual forms and connection forms. By assumption, we have

$$d\omega_j^i = - \sum_k \omega_k^i \wedge \omega_j^k.$$

Step 1. By Proposition 2, in a neighborhood of p there is a matrix $A = (A_j^i)$ of functions with $A(p)$ orthogonal and

$$(*) \quad dA = -\omega \wedge A.$$

Let $B = A^{-1}$ and define 1-forms $\phi^i = \sum_j B_j^i \theta^j$, which we can also write as

$$\theta = A \wedge \phi.$$

Step 2. We have

$$\begin{aligned} A \wedge d\phi &= d\theta - (dA \wedge \phi) \\ &= -\omega \wedge \theta + (\omega \wedge A \wedge \phi) && \text{by } (*) \text{ and the first structural equation} \\ &= -\omega \wedge A \wedge \phi + (\omega \wedge A \wedge \phi) && \text{by definition of } \theta \\ &= 0. \end{aligned}$$

So $d\phi = 0$, and consequently there are functions x^i with $\phi^i = dx^i$.

Step 3. We claim that x^1, \dots, x^n is the desired coordinate system, i.e., that the natural frame $\partial/\partial x^1, \dots, \partial/\partial x^n$ is orthonormal. To prove this, we first note that the 1-forms ϕ^i satisfy $\phi^i(Y_j) = B_j^i$, so

$$\frac{\partial}{\partial x^j} = \sum_i A_j^i Y_i.$$

Since the Y_j are orthonormal, it suffices to prove that (A_j^i) is always an orthogonal matrix. This is a consequence of $(*)$ and the fact that $\omega = (\omega_j^i)$ is skew-symmetric. The argument for this conclusion has already been given on page 36: if t denotes the transpose, we note that $A \cdot A^t$ satisfies the differential equation

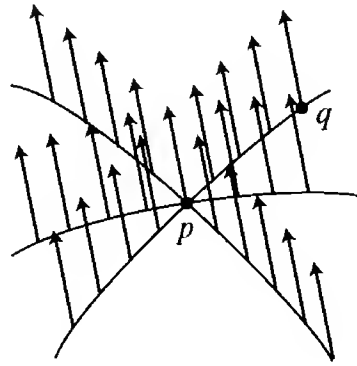
$$\begin{aligned} d(A \cdot A^t) &= -(\omega \wedge A \cdot A^t) - (A \cdot (\omega \wedge A)^t) && \text{by } (*) \\ &= -(\omega \wedge A \cdot A^t) - (A \cdot A^t \wedge \omega^t) \\ &= -\omega \wedge (A \cdot A^t) + (A \cdot A^t) \wedge \omega, \end{aligned}$$

with the initial condition $A \cdot A^t(0) = I$. By uniqueness, the solution is $A \cdot A^t = I$ everywhere. ♦

We leave it to the reader to correlate the three *Steps* in this proof with those in previous sections. Note that the first structural equation involves the symmetry of the connection, as shown by the proof of Theorem 5, and that skew-symmetry of ω is equivalent to the definition of the Christoffel symbols, since it is the condition which determines the ω_j^i in Proposition 4.

In contrast to previous chapters, in this one we are going to give more than one proof of the Test Case. Although the proof just given is a natural use of the structural equations as integrability conditions, it does not illustrate the fundamental principle to be used in the method of the repère mobile, which is *to choose the moving frame most suitable to the particular problem*. When we return to the study of surfaces in 3-space, or submanifolds of Riemannian manifolds in general, we will see many instances of this principle. In our present setup there is one especially important moving frame, the investigation of which will take some time.

Let X_{1p}, \dots, X_{np} be an orthonormal frame at M_p . In a sufficiently small neighborhood of p we can define a moving frame X_1, \dots, X_n by choosing $X_i(q)$ to be the parallel translate of X_{ip} along the unique geodesic from p to q . The



moving frame X_1, \dots, X_n is said to be **adapted** to the frame X_{1p}, \dots, X_{np} . In order to explore the properties of this moving frame, it is convenient to introduce the map

$$\Phi: \mathbb{R} \times M_p \rightarrow M$$

defined by

$$\Phi(t, X_p) = \exp(tX_p)$$

(actually, of course, Φ is usually not defined on all of $\mathbb{R} \times M_p$, but we will continue to write $\Phi: \mathbb{R} \times M_p \rightarrow M$ for convenience). On M_p we introduce the coordinate system t^1, \dots, t^n by $t^i(\sum_j a^j X_{jp}) = a^i$, and we use t for the standard coordinate system on \mathbb{R} ; we then let (t, t^1, \dots, t^n) denote the obvious

coordinate system on $\mathbb{R} \times M_p$. We now begin to describe the dual forms θ^i and the connection forms ω_j^i in terms of their pull-backs to $\mathbb{R} \times M_p$.

7. PROPOSITION. When we write $\Phi^*\theta^i$ and $\Phi^*\omega_j^i$ in terms of dt^1, \dots, dt^n and dt , we have

$$\begin{aligned}\Phi^*\theta^i &= t^i dt + \bar{\theta}^i \\ \Phi^*\omega_j^i &= \bar{\omega}_j^i,\end{aligned}$$

where $\bar{\theta}^i$ and $\bar{\omega}_j^i$ are 1-forms which do *not* involve dt .

PROOF. We can always write

$$\begin{aligned}\Phi^*\theta^i &= f_i dt + \bar{\theta}^i \\ \Phi^*\omega_j^i &= g_{ij} dt + \bar{\omega}_j^i;\end{aligned}$$

it is only necessary to identify the f_i and g_{ij} . To do this, we fix a^1, \dots, a^n and consider the geodesic

$$\gamma(s) = \exp\left(s \sum_i a^i X_{ip}\right), \quad s \in (-\varepsilon, \varepsilon).$$

This can be written as $\gamma = \Phi \circ c$, where $c: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R} \times M_p$ is $c(s) = (s, \sum_i a^i X_{ip})$. So

$$\begin{aligned}\gamma^*\theta^i(s) &= c^*\Phi^*\theta^i(s) = f_i(s, a^1, \dots, a^n) dt \\ \gamma^*\omega_j^i(s) &= c^*\Phi^*\omega_j^i(s) = g_{ij}(s, a^1, \dots, a^n) dt.\end{aligned}$$

On the other hand,

$$\gamma^*\theta^i \left(\frac{d}{dt} \Big|_s \right) = \theta^i \left(\frac{d\gamma}{dt} \Big|_s \right) = \theta^i \left(\sum_k a^k X_k(\gamma(s)) \right) = a^i,$$

which shows that $f_i = t^i$. Also

$$\gamma^*\omega_j^i \left(\frac{d}{dt} \Big|_s \right) = \omega_j^i \left(\sum_k a^k X_k(\gamma(s)) \right) = \sum_k \Gamma_{kj}^i a^k;$$

this sum vanishes, because X_j is parallel along γ , which means that

$$0 = \nabla_{\sum_k a^k X_k} X_j = \sum_k a^k \nabla_{X_k} X_j = \sum_k a^k \sum_i \Gamma_{kj}^i X_i. \quad \spadesuit$$

In our next result we will look at “partial derivatives” of the $\bar{\theta}^i$ and $\bar{\omega}_j^i$. If we are given an expression

$$\bar{\theta}^i = \sum_{j=1}^n g_j dt^j,$$

then we will use the symbol

$$\frac{\partial \bar{\theta}^i}{\partial t} \quad \text{for} \quad \sum_{j=1}^n \frac{\partial g_j}{\partial t} dt^j$$

and we will use similar symbols for the $\bar{\omega}_j^i$ [this definition depends on the particular coordinate system (t, t^1, \dots, t^n)]. Note that in

$$d\bar{\theta}^i = \sum_{j=1}^n \left(\frac{\partial g_j}{\partial t} dt + \sum_k \frac{\partial g_j}{\partial t^k} dt^k \right) \wedge dt^j,$$

the terms involving dt are precisely

$$dt \wedge \frac{\partial \bar{\theta}^i}{\partial t}.$$

8. PROPOSITION. We have the following “structural equations in polar coordinates”:

$$\begin{aligned} \frac{\partial \bar{\theta}^i}{\partial t} &= dt^i + \sum_k t^k \bar{\omega}_k^i & \bar{\theta}^i(0, X) &= 0 \quad \text{for all } X \in M_p \\ \frac{\partial \bar{\omega}_j^i}{\partial t} &= \sum_{k,l} (\mathbf{R}^i_{jkl} \circ \Phi) t^i \bar{\theta}^l & \bar{\omega}_j^i(0, X) &= 0 \quad \text{for all } X \in M_p \end{aligned}$$

PROOF. By Proposition 7 we have

$$\begin{aligned} (1) \quad \Phi^*(d\theta^i) &= d(\Phi^*\theta^i) = dt^i \wedge dt + dt \wedge \frac{\partial \bar{\theta}^i}{\partial t} + \text{terms not involving } dt \\ (2) \quad \Phi^*(d\omega_j^i) &= d(\Phi^*\omega_j^i) = dt \wedge \frac{\partial \bar{\omega}_j^i}{\partial t} + \text{terms not involving } dt. \end{aligned}$$

On the other hand, by the structural equations (Theorem 5) and Proposition 7 again, we have

$$(3) \quad \begin{aligned} \Phi^*(d\theta^i) &= - \sum_k \Phi^* \omega_k^i \wedge \Phi^* \theta^k \\ &= - \sum_k \bar{\omega}_k^i \wedge (t^k dt + \bar{\theta}^k) \end{aligned}$$

$$(4) \quad \begin{aligned} \Phi^*(d\omega_j^i) &= - \sum_k \Phi^* \omega_k^i \wedge \Phi^* \omega_j^k + \frac{1}{2} \Phi^* \left(\sum_{k,l} \mathbf{R}^i_{jkl} \theta^k \wedge \theta^l \right) \\ &= - \sum_k \bar{\omega}_k^i \wedge \bar{\omega}_j^k + \frac{1}{2} \sum_{k,l} (\mathbf{R}^i_{jkl} \circ \Phi) (t^k dt + \bar{\theta}^k) \wedge (t^l dt + \bar{\theta}^l). \end{aligned}$$

Comparing the coefficients of dt in (1) and (3) we obtain

$$dt^i - \frac{\partial \bar{\theta}^i}{\partial t} = - \sum_k t^k \bar{\omega}_k^i,$$

which gives us the first equation in the theorem. Similarly, from (2) and (4) we obtain

$$dt \wedge \frac{\partial \bar{\omega}_j^i}{\partial t} = \frac{1}{2} \sum_{k,l} (\mathbf{R}^i_{jkl} \circ \Phi) [dt \wedge (t^k \bar{\theta}^l - t^l \bar{\theta}^k)];$$

together with the relation $\mathbf{R}^i_{jkl} = -\mathbf{R}^i_{jlk}$, this gives the second equation.

Since $\Phi(0, X) = p$ for all $X \in M_p$, we have $\Phi_{*(0,X)} = 0$, which gives the “initial conditions”

$$\bar{\theta}^i(0, X) = 0, \quad \bar{\omega}_j^i(0, X) = 0. \quad \blacklozenge$$

9. COROLLARY. The forms $\bar{\theta}^i$ satisfy the second order differential equation

$$\frac{\partial^2 \bar{\theta}^i}{\partial t^2} = \sum_{j,k,l} (\mathbf{R}^i_{jkl} \circ \Phi) t^j t^k \bar{\theta}^l$$

with the initial conditions

$$\begin{aligned} \bar{\theta}^i(0, X) &= 0 \\ \frac{\partial \bar{\theta}^i}{\partial t}(0, X) &= dt^i. \end{aligned}$$

PROOF. Differentiate the first equation in Proposition 8 and substitute in from the second. \blacklozenge

It should not be hard to see that for all $X \in M_p$ we have

$$(*) \quad \Phi_* \left(\frac{\partial}{\partial t^j} \Big|_{(1, X)} \right) = \exp_* \left(\frac{\partial}{\partial t^j} \Big|_X \right);$$

in this equation $\partial/\partial t^j$ denotes a tangent vector in $\mathbb{R} \times M_p$ as well as one in M_p , since we are using the same symbol t^j for a coordinate function on $\mathbb{R} \times M_p$ as on M_p . We have also written Φ_* instead of $\Phi_*(1, X)$, etc. Equation (*) shows that θ is determined by \exp_* and $\bar{\theta}^i(1, X)$. This makes Corollary 9 particularly significant, since it determines $\bar{\theta}^i(1, X)$ in terms of the functions $\mathbf{R}^i_{jkl} \circ \Phi$. As a first illustration of this point, we bore ourselves to tears by twice again proving

10. THEOREM (THE TEST CASE; FIFTH VERSION). Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. *Step 1.* Let X_1, \dots, X_n be the moving frame adapted to an orthonormal frame X_{1p}, \dots, X_{np} for M_p . From Proposition 8 we have

$$\frac{\partial \bar{\omega}_j^i}{\partial t} = 0, \quad \bar{\omega}_j^i(0, X) = 0,$$

which shows that $\bar{\omega}_j^i = 0$. This implies that $\omega_j^i = 0$, since (*) shows that Φ_* is a diffeomorphism at $(1, X)$ and

$$\bar{\omega}_j^i = \Phi^* \omega_j^i = \omega_j^i \circ \Phi_*.$$

Step 2. Since $\omega_j^i = 0$, the first structural equation shows that $d\theta^i = 0$. So

$$\begin{aligned} 0 &= d\theta^i(X_j, X_k) = X_j(\theta^i(X_k)) - X_k(\theta^i(X_j)) - \theta^i([X_j, X_k]) \\ &= -\theta^i([X_j, X_k]). \end{aligned}$$

Thus $[X_j, X_k] = 0$ for all j, k , so there is a coordinate system x^1, \dots, x^n with $X_i = \partial/\partial x^i$.

Step 3. This is the desired coordinate system, since the X_i are obtained from the X_{ip} by parallel translation, and are consequently everywhere orthonormal. ♦

In this proof it is still possible to separate the argument into the standard three steps. But in the next proof everything happens at once.

11. THEOREM (THE TEST CASE; SIXTH VERSION). Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. Let X_1, \dots, X_n be the moving frame adapted to an orthonormal frame X_{1p}, \dots, X_{np} for M_p . From Corollary 9 we have

$$\frac{\partial^2 \bar{\theta}^i}{\partial t^2} = 0, \quad \bar{\theta}^i(0, X) = 0, \quad \frac{\partial \bar{\theta}^i}{\partial t}(0, X) = dt^i,$$

which implies that

$$\bar{\theta}^i(t, X) = t dt^i, \quad \text{in particular } \bar{\theta}^i(1, X) = dt^i.$$

So

$$\begin{aligned} \delta_j^i &= \bar{\theta}^i(1, X) \left(\frac{\partial}{\partial t^j} \Big|_{(1, X)} \right) = \Phi^* \theta^i \left(\frac{\partial}{\partial t^j} \Big|_{(1, X)} \right) \quad [\text{for convenience we do not write } \Phi^* \theta^i(1, X)] \\ &= \theta^i \left(\exp_* \left(\frac{\partial}{\partial t^j} \Big|_X \right) \right) \quad \text{by (*).} \end{aligned}$$

This shows that the value of the vector field X_j at $\exp X$ is

$$X_j(\exp X) = \exp_* \left(\frac{\partial}{\partial t^j} \Big|_X \right).$$

Now the vector fields $\partial/\partial t^j$ on M_p are orthonormal with respect to the usual Riemannian metric $\langle \cdot, \cdot \rangle$ on M_p , so this equation shows that $\exp: (M_p, \langle \cdot, \cdot \rangle) \rightarrow (M, \langle \cdot, \cdot \rangle)$ is an isometry. ♦

Readers may sort out for themselves the vestigial forms in which the three *Steps* appear in this last proof. One thing does seem worth pointing out explicitly. In the two closely related proofs of Theorems 10 and 11 we use the device of expressing the integrability conditions $R = 0$ in terms of the map Φ . This is roughly equivalent to the method outlined in Problem I.6-8, where we solve a system of partial differential equations in \mathbb{R}^n by reducing them to ordinary equations along lines through the origin.

The proof of Theorem 11, similar to the previous proof as it may be, is particularly important to us, for the methods used may be generalized to arbitrary Riemannian manifolds $(M, \langle \cdot, \cdot \rangle)$. To do this we introduce 1-forms $\bar{\theta}^i$ on M_p

by $\bar{\bar{\theta}}^i(X) = \bar{\theta}^i(1, X)$. More precisely, if $X \in M_p$ and $v_X \in (M_p)_X$, then we have a tangent vector

$$(0, v_X) \in (\mathbb{R} \times M_p)_{(1, X)}$$

(recall that the tangent space of the product of two manifolds is isomorphic to the direct sum of the tangent spaces of the manifolds), so we may define

$$\bar{\bar{\theta}}^i(v_X) \quad [= \bar{\bar{\theta}}^i(X)(v_X)] \quad = \bar{\theta}^i(1, X)((0, v_X)).$$

In particular, we have (leaving out the arguments for $\bar{\theta}^i$ and $\bar{\bar{\theta}}^i$)

$$(**) \quad \bar{\bar{\theta}}^i \left(\frac{\partial}{\partial t^j} \Big|_X \right) = \bar{\theta}^i \left(\frac{\partial}{\partial t^j} \Big|_{(1, X)} \right).$$

Now define a tensor $\langle \ , \ \rangle$ of type $\binom{2}{0}$ on M_p by

$$\langle \ , \ \rangle = \sum_{i=1}^n \bar{\bar{\theta}}^i \otimes \bar{\theta}^i.$$

12. THEOREM. The map $\exp: (M_p, \langle \ , \ \rangle) \rightarrow (M, \langle \ , \ \rangle)$ is an isometry (in a neighborhood of $0 \in M_p$ on which \exp is a diffeomorphism).

PROOF. Since

$$\left\langle \sum_{j=1}^n a^j X_j, \sum_{j=1}^n b^j X_j \right\rangle = \sum_{i=1}^n a^i b^i = \sum_{i=1}^n \theta^i \left(\sum_{j=1}^n a^j X_j \right) \cdot \theta^i \left(\sum_{j=1}^n b^j X_j \right),$$

we have $\langle \ , \ \rangle = \sum_i \theta^i \otimes \theta^i$. On the other hand, we also have

$$\begin{aligned} \left\langle \frac{\partial}{\partial t^j} \Big|_X, \frac{\partial}{\partial t^k} \Big|_X \right\rangle &= \sum_{i=1}^n \bar{\bar{\theta}}^i \left(\frac{\partial}{\partial t^j} \Big|_X \right) \bar{\theta}^i \left(\frac{\partial}{\partial t^k} \Big|_X \right) \\ &= \sum_{i=1}^n \bar{\theta}^i \left(\frac{\partial}{\partial t^j} \Big|_{(1, X)} \right) \bar{\theta}^i \left(\frac{\partial}{\partial t^k} \Big|_{(1, X)} \right) \quad \text{by } (**) \\ &= \sum_{i=1}^n \theta^i \left(\exp_* \left(\frac{\partial}{\partial t^j} \Big|_X \right) \right) \cdot \theta^i \left(\exp_* \left(\frac{\partial}{\partial t^k} \Big|_X \right) \right), \end{aligned}$$

which means that

$$\langle \ , \ \rangle = \exp^* \left(\sum_i \theta^i \otimes \theta^i \right) = \exp^* \langle \ , \ \rangle. \quad \blacklozenge$$

Recall (page 194) that for a 2-dimensional subspace $W \subset M_q$ of the tangent space of a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ we define the sectional curvature $K(W)$ as

$$K(W) = \frac{\langle R(A, B)B, A \rangle}{\|A, B\|^2}, \quad A, B \text{ a basis for } W.$$

Let $p \in M$ be some fixed point. For every $X \in M_p$ and 2-dimensional subspace $V \subset M_p$, we will let $L(X, V)$ be the sectional curvature $K(W)$, where $W \subset M_{\exp X}$ is the parallel translate of V along the geodesic $t \mapsto \exp tX$.

13. COROLLARY (THE CURVATURE DETERMINES THE METRIC).

Let M and M' be two Riemannian manifolds, and $T: M_p \rightarrow M'_{p'}$, an isometry for some $p \in M$ and $p' \in M'$. Suppose that $L(X, V) = L'(T(X), T(V))$ for all 2-dimensional subspaces $V \subset M_p$ and all sufficiently small X . Then there is an isometry from some neighborhood of $p \in M$ to a neighborhood of $p' \in M'$.

PROOF. Choose an orthonormal frame $X_{1p}, \dots, X_{np} \in M_p$, let X_1, \dots, X_n be the adapted moving frame in M , and let X'_1, \dots, X'_n be the moving frame in M' adapted to $T(X_{1p}), \dots, T(X_{np})$. Let $\Phi: \mathbb{R} \times M_p \rightarrow M$ and $\Phi': \mathbb{R} \times M'_{p'} \rightarrow M'$ be as defined previously, let $\bar{\theta}^i$ and $\bar{\theta}'^i$ be the corresponding forms on $\mathbb{R} \times M_p$ and $\mathbb{R} \times M'_{p'}$, and let $S: \mathbb{R} \times M_p \rightarrow \mathbb{R} \times M'_{p'}$ be $S(a, X) = (a, T(X))$. From the definition of \mathbf{R}^i_{jkl} before Theorem 5 we see that

$$\mathbf{R}^i_{jkl} = \langle R(X_k, X_l)X_j, X_i \rangle.$$

The hypotheses of the theorem therefore imply that

$$\mathbf{R}^i_{jij} \circ \Phi = \mathbf{R}'^i_{jij} \circ (\Phi' \circ S) \quad \text{for all } i, j.$$

From Proposition 4-12 we deduce that

$$\mathbf{R}^i_{jkl} \circ \Phi = \mathbf{R}'^i_{jkl} \circ (\Phi' \circ S) \quad \text{for all } i, j, k, l.$$

Now Corollary 9 (and the uniqueness of solutions of differential equations with given initial conditions) implies that

$$\bar{\theta}^i = \bar{\theta}'^i \circ S,$$

and therefore that

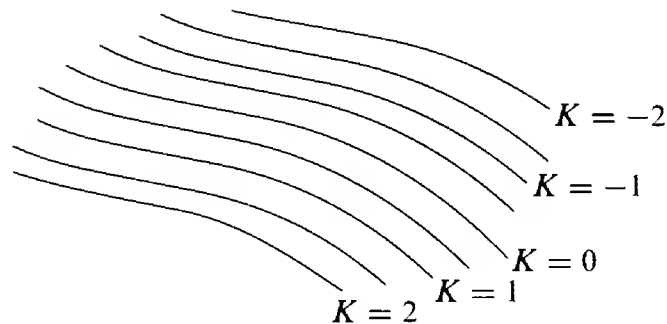
$$\bar{\bar{\theta}}^i = \bar{\bar{\theta}}'^i \circ T.$$

This means that T is an isometry of $(M_p, \langle \cdot, \cdot \rangle)$ and $(M'_{p'}, \langle \cdot, \cdot \rangle')$. The result then follows from Theorem 12. ♦

This corollary is the main assertion made by Riemann in his Habilitation lecture. We have proved a purely local result, but global results have also been obtained; see Ambrose, *Parallel translation of Riemannian curvature*, Annals of Math. **64** (1956), 337–363.

The result of Corollary 13 is perhaps not what the reader may have understood by the assertion that “the curvature determines the metric”, for it involves the parallel translation in the manifold, and not merely the curvature. Notice, however, that any rigorous statement about curvature determining the metric must involve a map $f: M \rightarrow M'$, so that we know what it means to compare curvature in M with curvature in M' ; in our case the map f is $\exp_{p'} \circ (\exp_p)^{-1}$. In this connection the following rather different question has always seemed to me the more interesting one. Suppose we have a diffeomorphism $f: M \rightarrow M'$ such that for every 2-dimensional $V \in M_q$ we have $K(V) = K'(f_*(V))$; then is f an isometry? It is easy to see that as stated this is *not* true, because *any* diffeomorphism $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies the hypothesis (when \mathbb{R}^n has its usual Riemannian metric), but not necessarily the conclusion. It is also easy to obtain other examples. Consider the sphere $S^n \subset \mathbb{R}^{n+1}$ of radius a , with the induced Riemannian metric. For $n = 2$, we know that $K(S^2_p) = K(p) = 1/a^2$. If $n > 2$ and \mathcal{O} is a neighborhood of 0 in a 2-dimensional subspace $V \subset M_p$, then $\exp(\mathcal{O})$ is isometric to a portion of S^2 , so we have $K(V) = 1/a^2$ for all such V . Consequently, once again any diffeomorphism $f: S^n \rightarrow S^n$ satisfies the hypothesis of our question, but not necessarily the conclusion. As we shall see in Addendum 2, there are also examples where all $K(V)$ are the same negative number.

In addition to these rather special counterexamples, it is simple to construct infinitely many other 2-dimensional examples. If we have a 2-dimensional manifold M such that the sets $K = \text{constant}$ give a foliation of M , then we can choose



$f: M \rightarrow M$ to be *any* diffeomorphism which keeps each folium fixed as a set. There are also specific classical examples of 2-manifolds M, M' which are not isometric under any map, but for which there is a diffeomorphism $f: M \rightarrow M'$

with $K'(f(p)) = K(p)$ [see pg. III.166]. Perhaps these examples are responsible for the fact that the higher dimensional cases remained unsettled for so long. Though it might be natural to assume that this case is even more hopeless, just the opposite is true: when $n > 2$, a diffeomorphism which preserves sectional curvatures is, *roughly speaking*, an isometry, except for the special counterexamples mentioned in the previous paragraph. For details the reader is referred to R. S. Kulkarni, *Curvature and Metric*, Annals of Math. **91** (1970), 311–331.

Before we proceed further with the study of moving frames, we pause to note that in the 2-dimensional case, Corollary 9 (and Corollary 13 which depends on it) have really been available to us since Chapter 3. Recall that we chose polar coordinates (ρ, ϕ) on M_p , and used them to introduce polar coordinates $(r, \varphi) = (\rho, \phi) \circ \exp^{-1}$ on a neighborhood of p [see page 136]. If we write

$$\langle \cdot, \cdot \rangle = dr \otimes dr + G d\phi \otimes d\phi,$$

and define g on \mathbb{R}^2 by

$$g = G \circ \exp \circ (\rho, \phi)^{-1},$$

then we have the following formulas (collected together on page 145):

$$\begin{aligned} \sqrt{g}(0, \phi) &= 0 \\ \frac{\partial \sqrt{g}}{\partial \rho}(0, \phi) &= 1 \\ \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) &= -\sqrt{g}(\rho, \phi) \cdot K(\exp(\rho, \phi)). \end{aligned}$$

These equations are easily seen to be equivalent to the equations in Corollary 9, and can be used the same way. For example, if $K = 0$ in a neighborhood of p , then

$$\frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) = 0;$$

together with the initial conditions, this shows that $\sqrt{g} = \rho$, so

$$\langle \cdot, \cdot \rangle = dr \otimes dr + r^2 d\phi \otimes d\phi,$$

which is exactly the expression in polar coordinates for the usual Riemannian metric on \mathbb{R}^2 . Even the n -dimensional Test Case could be proved in this way, by considering $\exp(V)$ for various 2-dimensional $V \subset M_p$. However, the equations of Corollary 9 are generally most useful in higher dimensions. As an exercise in using them, Addendum 1 derives the form of the metric in an n -dimensional manifold of constant curvature.

Everything which we have done so far in this chapter has involved the Levi-Civita connection associated to a metric $\langle \cdot, \cdot \rangle$; moreover, we immediately interpreted the connection and curvature forms in terms of concepts defined in previous chapters. But the method of the repère mobile is meant to treat arbitrary connections, and was used to define the curvature tensor before the ∇ operator had been invented. For the remainder of this chapter we will be concerned with this independent development of the theory of connections. We want to describe any connection in terms of “connection forms” ω_j^i , but we do not necessarily want the matrix $\omega = (\omega_j^i)$ to be skew-symmetric, so we do not have Proposition 4 to guide us. We do want to have the equations

$$\nabla_{X_k} X_j = \sum_i \omega_j^i(X_k) X_i,$$

which are a consequence of Theorem 5; these equations can also be written

$$\nabla X_j = \sum_i \omega_j^i \cdot X_i.$$

For convenience, we let $\mathbf{X} = (X_1, \dots, X_n)$ and abbreviate this equation as

$$\nabla \mathbf{X} = \mathbf{X} \cdot \omega \quad (\text{recall the notation introduced on page 261ff}).$$

Now consider another moving frame $\mathbf{X}' = \mathbf{X} \cdot a$. We want to have

$$\begin{aligned} \sum_i \omega_j'^i X'_i &= \nabla X'_j = \nabla \left(\sum_l a_j^l X_l \right) \\ &= \sum_l da_j^l X_l + \sum_l a_j^l \nabla X_l \quad \begin{array}{l} [\text{recall the formula for } \nabla_X fY \\ \text{and note that } X(f) = df(X)]. \end{array} \end{aligned}$$

Using our matrix notation, this means that we want

$$\begin{aligned} \mathbf{X}' \cdot \omega' &= \nabla \mathbf{X}' = \nabla(\mathbf{X} \cdot a) \\ &= \mathbf{X} \cdot da + \nabla \mathbf{X} \cdot a \\ &= \mathbf{X} \cdot da + \mathbf{X} \cdot (\omega \cdot a), \end{aligned}$$

so we want

$$\mathbf{X} \cdot (a\omega') = (\mathbf{X} \cdot a)\omega' = \mathbf{X}' \cdot \omega' = \mathbf{X} \cdot (da + \omega a)$$

for all moving frames \mathbf{X} . Thus we want the condition

$$(*) \quad \omega' = a^{-1} da + a^{-1} \omega a.$$

Hence we are led to the following definition:*

A **(Cartan) connection** on a manifold M is an assignment of a matrix $\omega = (\omega_j^i)$ of 1-forms to every moving frame \mathbf{X} such that equation (*) holds between the 1-forms ω_j^i assigned to the moving frame \mathbf{X} and the 1-forms $\omega_j'^i$ assigned to the moving frame $\mathbf{X}' = \mathbf{X} \cdot a$.

Given a Cartan connection, and a moving frame $\mathbf{X} = X_1, \dots, X_n$, the 1-forms ω_j^i which are assigned to \mathbf{X} are called the **connection forms** for \mathbf{X} , and we define the **dual forms** θ^i by $\theta^i(X_j) = \delta_j^i$. From these forms we can define all the tensors which arise in Chapters 5 and 6. What follows is an outline of such a development of the theory of Cartan connections, independently of previous considerations.

We begin with a simple observation about the consistency of the transformation laws (*). Suppose $\mathbf{X}'' = \mathbf{X}' \cdot b = (\mathbf{X} \cdot a) \cdot b = \mathbf{X} \cdot (ab)$ is another moving frame, and that ω', ω'' satisfy

$$\begin{aligned}\omega' &= a^{-1} da + a^{-1} \omega a \\ \omega'' &= b^{-1} db + b^{-1} \omega' b.\end{aligned}$$

Then

$$\begin{aligned}\omega'' &= b^{-1} db + b^{-1} (a^{-1} da + a^{-1} \omega a) b \\ &= [b^{-1} a^{-1} (da) b + b^{-1} a^{-1} a (db)] + b^{-1} a^{-1} \omega ab \\ &= (ab)^{-1} d(ab) + (ab)^{-1} \omega ab.\end{aligned}$$

This shows that if we are given connection forms for a certain set of moving frames whose domains cover M , and the various pairs of connection forms all satisfy (*), then there is a unique Cartan connection that assigns these forms to these particular moving frames. In view of this remark, it is very easy to determine a connection from a Riemannian metric.

14. PROPOSITION. On a Riemannian manifold $(M, \langle \ , \ \rangle)$ there is a unique Cartan connection with the property that the connection forms ω_j^i for any

*We are using the term "Cartan connection" as a convenient label, but the reader should be warned that in the literature this term is used for a different concept.

orthonormal moving frame X satisfy

$$\begin{aligned}\omega_j^i &= -\omega_i^j \\ d\theta^i &= \sum_k \theta^k \wedge \omega_k^i.\end{aligned}$$

PROOF. We already know, by Proposition 4, that for any moving frame there are unique 1-forms ω_j^i with this property. We just have to check that if \mathbf{X} and $\mathbf{X}' = \mathbf{X} \cdot a$ are orthonormal moving frames, then

$$\omega' = a^{-1} da + a^{-1} \omega a.$$

In view of uniqueness, we just have to show that if the forms ω_j^i satisfy the conditions of the theorem, and the forms ω'^i_j are defined by this formula, then they satisfy the same conditions,

$$\begin{aligned}\text{(a)} \quad \omega'^i_j &= -\omega'^j_i \\ \text{(b)} \quad d\theta'^i &= -\sum_k \omega'^i_k \wedge \theta'^k.\end{aligned}$$

Since \mathbf{X} and \mathbf{X}' are orthonormal, the matrix a is everywhere orthogonal, $a \cdot a^t = I$, where t denotes the transpose. So if 0 denotes the zero matrix we have

$$0 = da \cdot a^t + a \cdot da^t,$$

or

$$da^t = -a^{-1} \cdot da \cdot a^t = -a^t \cdot da \cdot a^t.$$

Consequently,

$$\begin{aligned}\text{(1)} \quad (a^{-1} da)^t &= (a^t da)^t = da^t \cdot a = -a^t \cdot da \cdot a^t \cdot a = -a^t \cdot da = -a^{-1} da \\ \text{(2)} \quad (a^{-1} \omega a)^t &= (a^t \omega a)^t = a^t \omega^t a = -a^t \omega a = -a^{-1} \omega a.\end{aligned}$$

Clearly (1) and (2) imply (a).

To prove (b), we first note that

$$\theta'^i(X_j) = \theta'^i \left(\sum_k (a^{-1})^k_j X'_k \right) = (a^{-1})^i_j = \sum_k (a^{-1})^i_k \theta^k(X_j),$$

so $\theta'^i = \sum_k (a^{-1})^i_k \theta^k$, or

$$\theta' = a^{-1} \cdot \theta.$$

Consequently,*

$$(3) \quad \begin{aligned} d\theta' &= (-a^{-1} \cdot da \cdot a^{-1}) \wedge \theta + a^{-1} d\theta \\ &= -a^{-1} da \wedge a^{-1} \theta - (a^{-1} \omega \wedge \theta). \end{aligned}$$

On the other hand,

$$(4) \quad \begin{aligned} \omega' \wedge \theta' &= (a^{-1} da + a^{-1} \omega a) \wedge (a^{-1} \theta) \\ &= a^{-1} da \wedge a^{-1} \theta + (a^{-1} \omega \wedge \theta). \end{aligned}$$

Clearly (3) and (4) imply (b). ♦

When we pass from the particular connection of Proposition 14 to a general Cartan connection, both structural equations need correction terms. If ω_j^i are the connection forms for a moving frame \mathbf{X} , with dual forms θ^i , we define 2-forms Θ^i and Ω_j^i by

$$\begin{aligned} d\theta &= -\omega \wedge \theta + \Theta & \text{i.e.,} & \quad d\theta^i = -\sum_k \omega_k^i \wedge \theta^k + \Theta^i \\ d\omega &= -\omega \wedge \omega + \Omega & \text{i.e.,} & \quad d\omega_j^i = -\sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i. \end{aligned}$$

We call the Θ^i and Ω_j^i the **torsion forms** and **curvature forms** for the moving frame X . We now compute the transformation formulas for these forms.

15. PROPOSITION. If Θ'^i and $\Omega_j'^i$ are the torsion and connection forms for another moving frame $\mathbf{X}' = \mathbf{X} \cdot a$, then

$$\begin{aligned} \Theta' &= a^{-1} \cdot \Theta \\ \Omega' &= a^{-1} \Omega a. \end{aligned}$$

*To compute $d(a^{-1})$ we differentiate $a \cdot a^{-1} = I$ to obtain

$$\begin{aligned} da \cdot a^{-1} + a \cdot d(a^{-1}) &= 0, \\ d(a^{-1}) &= -a^{-1} \cdot da \cdot a^{-1}. \end{aligned}$$

PROOF. We have

$$\begin{aligned}
 (1) \quad & \omega' = a^{-1} da + a^{-1} \omega a \\
 (2) \quad & d\theta = -\omega \wedge \theta + \Theta \\
 (3) \quad & d\theta' = -\omega' \wedge \theta' + \Theta' \\
 (4) \quad & d\omega = -\omega \wedge \omega + \Omega \\
 (5) \quad & d\omega' = -\omega' \wedge \omega' + \Omega'
 \end{aligned}$$

and, as in the proof of Proposition 14,

$$\begin{aligned}
 (6) \quad & \theta' = a^{-1} \theta \\
 (7) \quad & d\theta' = -a^{-1} da \wedge a^{-1} \theta + a^{-1} d\theta.
 \end{aligned}$$

So

$$(8) \quad d\theta' = -a^{-1} da \wedge a^{-1} \theta + a^{-1} (-\omega \wedge \theta + \Theta) \quad \text{by (2), (7).}$$

We also have

$$\begin{aligned}
 (9) \quad d\theta' &= -\omega' \wedge \theta' + \Theta' && \text{by (3)} \\
 &= -(a^{-1} da + a^{-1} \omega a) \wedge a^{-1} \theta + \Theta' && \text{by (1), (6)} \\
 &= -a^{-1} da \wedge a^{-1} \theta - a^{-1} (\omega \wedge \theta) + \Theta'.
 \end{aligned}$$

Comparison of (8) and (9) gives $a^{-1} \Theta = \Theta'$.

From (1) we obtain

$$\begin{aligned}
 (10) \quad d\omega' &= [-a^{-1} da a^{-1} \wedge da] + (-a^{-1} da a^{-1}) \wedge \omega a \\
 &\quad + a^{-1} d\omega a - a^{-1} \omega \wedge da \\
 &= [-a^{-1} da a^{-1} \wedge da] + (-a^{-1} da a^{-1}) \wedge \omega a \\
 &\quad + a^{-1} (-\omega \wedge \omega + \Omega) a - a^{-1} \omega \wedge da \quad \text{by (4).}
 \end{aligned}$$

We also have

$$\begin{aligned}
 (11) \quad d\omega' &= -\omega' \wedge \omega' + \Omega' \quad \text{by (5)} \\
 &= -(a^{-1} da + a^{-1} \omega a) \wedge (a^{-1} da + a^{-1} \omega a) + \Omega' \quad \text{by (1)} \\
 &= -(a^{-1} da \wedge a^{-1} da) - (a^{-1} da a^{-1} \wedge \omega a) - (a^{-1} \omega \wedge da) \\
 &\quad - (a^{-1} \omega \wedge \omega a) + \Omega'.
 \end{aligned}$$

Comparison of (10) and (11) gives $a^{-1} \Omega a = \Omega'$. ♦

The three relations

$$(1) \omega' = a^{-1} da + a^{-1} \omega a$$

$$(2) \Theta' = a^{-1} \Theta$$

$$(3) \Omega' = a^{-1} \Omega a$$

are precisely what enable us to define

$$(1) \nabla Y \text{ for vector fields } Y$$

$$(2) T(X, Y) \text{ for tangent vectors } X, Y$$

$$(3) R(X, Y)Z \text{ for tangent vectors } X, Y, Z.$$

We essentially know this already for the ∇ operator, which we consider first. Given a moving frame $\mathbf{X} = (X_1, \dots, X_n)$ with connection forms ω_j^i , we define

$$\nabla \mathbf{X} = \mathbf{X} \cdot \omega \quad [\text{i.e., } \nabla X_j = \sum_i \omega_j^i X_i \quad \text{or} \quad \nabla_X X_j = \sum_i \omega_j^i(X) X_i]$$

and extend this to arbitrary vector fields $Y = \sum_j b^j X_j$ by defining

$$\nabla \cdot \left(\sum_j b^j X_j \right) = \sum_j db^j \cdot X_j + b^j \nabla X_j.$$

For another moving frame $\mathbf{X}' = \mathbf{X} \cdot a$ with connection forms $\omega_j'^i$ we have

$$\begin{aligned} \nabla \mathbf{X}' &= \mathbf{X}' \cdot \omega' = (\mathbf{X} \cdot a) \cdot \omega' = (\mathbf{X} \cdot a) \cdot [a^{-1} da + a^{-1} \omega a] \\ &= \mathbf{X} \cdot [a \cdot (a^{-1} da + a^{-1} \omega a)] \\ &= \mathbf{X} \cdot da + \mathbf{X} \cdot (\omega \cdot a) = \mathbf{X} \cdot da + (\mathbf{X} \cdot \omega) \cdot a \\ &= \mathbf{X} \cdot da + \nabla \mathbf{X} \cdot a, \end{aligned}$$

which shows that the definition in terms of the two moving frames are consistent [this equation becomes

$$\begin{aligned} \nabla \left(\sum_j a_j^i X_j \right) &= \sum_i da_j^i \cdot X_i + \sum_k \left(\sum_i \omega_i^k a_j^i \right) X_k \\ &= \sum_i da_j^i X_i + \sum_i a_j^i \nabla X_i. \end{aligned}$$

when written out].

We next define

$$T(X_j, X_k) = \sum_i \Theta^i(X_j, X_k) \cdot X_i$$

and extend to arbitrary tangent vectors $X = \sum_j b^j X_j, Y = \sum_k c^k X_k$ by linearity,

$$T(X, Y) = \sum_{j,k} b^j c^k T(X_j, X_k).$$

For the moving frame $\mathbf{X}' = \mathbf{X} \cdot a$ we have

$$\begin{aligned} T(X'_\mu, X'_\nu) &= \sum_\rho \Theta'^\rho(X'_\mu, X'_\nu) X'_\rho \\ &= \sum_\rho \sum_l (a^{-1})^\rho_l \Theta^l \left(\sum_j a^j_\mu X_j, \sum_k a^k_\nu X_k \right) \sum_i a^i_\rho X_i \\ &= \sum_{j,k} a^j_\mu a^k_\nu \sum_i \Theta^i(X_j, X_k) X_i \\ &= \sum_{j,k} a^j_\mu a^k_\nu T(X_j, X_k), \end{aligned}$$

so the definitions in terms of the two moving frames are consistent. The tensor T is clearly alternating, since the Θ^i are 2-forms.

Finally, we define

$$R(X_k, X_l) X_j = \sum_i \Omega^i_j(X_k, X_l) X_i$$

and extend by linearity. For the moving frame $\mathbf{X}' = \mathbf{X} \cdot a$ we have

$$\begin{aligned} R(X'_\mu, X'_\nu) X'_\lambda &= \sum_\rho \Omega'^\rho_\lambda(X'_\mu, X'_\nu) X'_\rho \\ &= \sum_\rho \sum_{j,m} (a^{-1})^\rho_m \Omega^m_j a^j_\lambda \left(\sum_k a^k_\mu X_k, \sum_l a^l_\nu X_l \right) \sum_i a^i_\rho X_i \\ &= \sum_{j,k,l} a^k_\mu a^l_\nu a^j_\lambda \Omega^i_j(X_k, X_l) X_i, \end{aligned}$$

which again shows consistency. Clearly R is skew-symmetric in the first two arguments, since the Ω^i_j are 2-forms.

If we now define

$$\begin{aligned}\nabla_{X_k} X_j &= \sum_{i=1}^n \Gamma_{kj}^i X_i \\ T(X_j, X_k) &= \sum_{i=1}^n \mathbf{T}_{jk}^i X_i \\ R(X_k, X_l) X_j &= \sum_{i=1}^n \mathbf{R}_{jkl}^i X_i,\end{aligned}$$

then we have

$$\begin{aligned}\Gamma_{jk}^i &= \omega_j^i(X_k) \quad \text{or} \quad \omega_j^i = \sum_k \Gamma_{kj}^i \theta^k \\ \mathbf{T}_{jk}^i &= \Theta^i(X_j, X_k) \quad \text{or} \quad \Theta^i = \frac{1}{2} \sum_{j,k} \mathbf{T}_{jk}^i \theta^j \wedge \theta^k \\ \mathbf{R}_{jkl}^i &= \Omega_j^i(X_k, X_l) \quad \text{or} \quad \Omega_j^i = \frac{1}{2} \sum_{k,l} \mathbf{R}_{jkl}^i \theta^k \wedge \theta^l.\end{aligned}$$

Consequently, the “structural equations”

$$\begin{aligned}d\theta^i &= - \sum_k \omega_k^i \wedge \theta^k + \Theta^i = - \sum_k \omega_k^i \wedge \theta^k + \frac{1}{2} \sum_{j,k} \mathbf{T}_{jk}^i \theta^j \wedge \theta^k \\ d\omega_j^i &= - \sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i = - \sum_k \omega_k^i \wedge \omega_j^k + \frac{1}{2} \sum_{k,l} \mathbf{R}_{jkl}^i \theta^k \wedge \theta^l \\ \omega_j^i &= \sum_k \Gamma_{kj}^i \theta^k\end{aligned}$$

are purely a matter of definition. However, one can now easily reverse the computations in the proof of Theorem 5 to show that

$$\begin{aligned}T(X, Y) &= \nabla_X Y - \nabla_Y X - [X, Y] \\ R(X, Y) Z &= \nabla_X (\nabla_Y Z) - \nabla_Y (\nabla_X Z) - \nabla_{[X, Y]} Z,\end{aligned}$$

thus verifying that T and R are the torsion and curvature tensors for ∇ as defined previously. We leave this computation to the reader, as well as the task of deriving the general form of the structural equations in polar coordinates.

Even though the structural equations are merely definitions in this approach, we can derive new relations from them.

16. THEOREM. We have the following relations between θ , ω , Θ , and Ω :

(1) (Bianchi's first identity)

$$d\Theta + \omega \wedge \Theta = \Omega \wedge \theta$$

(2) (Bianchi's second identity)

$$d\Omega + (\omega \wedge \Omega) - (\Omega \wedge \omega) = 0.$$

(Notice that $\omega \wedge \Omega$ and $\Omega \wedge \omega$ are not equal up to sign, because Ω and ω are both matrices of forms and the order of matrix multiplication plays a role.)

PROOF. We have

$$\begin{aligned} 0 &= d(d\theta) = d(-\omega \wedge \theta + \Theta) \\ &= -d\omega \wedge \theta + (\omega \wedge d\theta) + d\Theta \\ &= -(-\omega \wedge \omega + \Omega) \wedge \theta + [\omega \wedge (-\omega \wedge \theta + \Theta)] + d\Theta \\ &= -\Omega \wedge \theta + \omega \wedge \Theta + d\Theta. \end{aligned}$$

Similarly,

$$\begin{aligned} 0 &= d(d\omega) = d(-\omega \wedge \omega + \Omega) \\ &= -d\omega \wedge \omega + (\omega \wedge d\omega) + d\Omega \\ &= -(-\omega \wedge \omega + \Omega) \wedge \omega + [\omega \wedge (-\omega \wedge \omega + \Omega)] + d\Omega \\ &= -\Omega \wedge \omega + (\omega \wedge \Omega) + d\Omega. \quad \spadesuit \end{aligned}$$

It takes quite a bit of calculation to convince oneself that the equations in Theorem 16 really are the Bianchi identities. This calculation is left to those readers who have more endurance than the author; we will merely consider the special case of the first Bianchi identity for a connection without torsion. In this case $\Theta = 0$, so the identity is just $\Omega \wedge \theta = 0$. Thus we have

$$0 = (\Omega \wedge \theta)^i = \sum_j \Omega_j^i \wedge \theta^j = \frac{1}{2} \sum_{j,k,l} \mathbf{R}^i_{jkl} \theta^k \wedge \theta^l \wedge \theta^j.$$

Applying this to (X_k, X_l, X_j) we obtain the familiar formula

$$\begin{aligned} 0 &= \frac{1}{2} \{ \mathbf{R}^i_{jkl} - \mathbf{R}^i_{jlk} + \mathbf{R}^i_{ljk} - \mathbf{R}^i_{lkj} + \mathbf{R}^i_{klj} - \mathbf{R}^i_{kjl} \} \\ &= \mathbf{R}^i_{jkl} + \mathbf{R}^i_{klj} + \mathbf{R}^i_{ljk}, \end{aligned}$$

(and thereby see why cyclic permutations of the indices should be involved). It turns out that taking the exterior derivative of the Bianchi identities does not give us any new relations, which indicates that these identities are the only general ones we should expect to find.

To complete the present chapter we also derive the relations satisfied by the curvature tensor for the Levi-Civita connection.

17. THEOREM. Let $(M, \langle \cdot, \cdot \rangle)$ be a Riemannian manifold, and consider the unique Cartan connection of Proposition 14. Then for every orthonormal moving frame we have

$$(1) \quad \Omega_j^i = -\Omega_i^j.$$

Consequently, the curvature tensor R satisfies

$$(2) \quad \langle R(X, Y)Z, W \rangle = -\langle R(X, Y)W, Z \rangle$$

$$(3) \quad \langle R(X, Y)Z, W \rangle = \langle R(Z, W)X, Y \rangle.$$

PROOF. Equation (1) is immediate from the fact that $\omega_j^i = -\omega_i^j$ (by assumption) and the second structural equation,

$$d\omega_j^i = -\sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i.$$

Since

$$\begin{aligned} \langle R(X_k, X_l)X_j, X_i \rangle &= \left\langle \sum_{\mu} \mathbf{R}^{\mu}_{jkl} X_{\mu}, X_i \right\rangle \\ &= \mathbf{R}^i_{jkl} = \Omega_j^i(X_k, X_l), \end{aligned}$$

equation (1) implies (2). Then, as before, (3) follows from Proposition 4-11. ❖

ADDENDUM 1

MANIFOLDS OF CONSTANT CURVATURE

A Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ has **constant curvature** K_0 if for all $p \in M$, and all 2-dimensional $V \subset M_p$ we have $K(V) = K_0$. If M is 2-dimensional, this just means that $K(p) = K_0$ for all $p \in M$. In this case, we can easily find the form of the metric $\langle \cdot, \cdot \rangle$ from the considerations on page 145. We use geodesic polar coordinates $(r, \varphi) = (\rho, \phi) \circ \exp^{-1}$ so that the metric is

$$\langle \cdot, \cdot \rangle = dr \otimes dr + G d\varphi \otimes d\varphi.$$

If g on \mathbb{R}^2 is defined by $g = G \circ \exp \circ (\rho, \phi)^{-1}$, then we have seen that

$$\begin{aligned}\sqrt{g}(0, \phi) &= 0 \\ \frac{\partial \sqrt{g}}{\partial \rho}(0, \phi) &= 1 \\ \frac{\partial^2 \sqrt{g}}{\partial \rho^2}(\rho, \phi) &= -K_0 \sqrt{g}(\rho, \phi).\end{aligned}$$

The general solution of the last equation is

$$\begin{aligned}\sqrt{g}(\rho, \phi) &= c_1 \sin \sqrt{K_0} \rho + c_2 \cos \sqrt{K_0} \rho & K_0 > 0 \\ \sqrt{g}(\rho, \phi) &= c_1 \sinh \sqrt{-K_0} \rho + c_2 \cosh \sqrt{-K_0} \rho & K_0 < 0.\end{aligned}$$

Taking into account the initial conditions, we find that

$$\sqrt{g} = \frac{s(\sqrt{|K_0|} \rho)}{\sqrt{|K_0|}},$$

where

$$s \text{ denotes } \begin{cases} \sin & \text{whenever we are dealing with } K_0 > 0 \\ \sinh & \text{''} & K_0 < 0. \end{cases}$$

It follows that the metric is given by

$$\langle \cdot, \cdot \rangle = dr \otimes dr + \frac{s^2(\sqrt{|K_0|} r)}{\sqrt{|K_0|}} d\varphi \otimes d\varphi.$$

To obtain the analogous results in the n -dimensional case, we will return to the equations of Corollary 9. However, we will need a preliminary result,

which applies to spaces of constant curvature as a special case. A point p in an n -dimensional Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ is called **isotropic** if the sectional curvatures $K(V)$ for all 2-dimensional $V \subset M_p$ have the same value K_0 ; this means that for all $X, Y \in M_p$ we have

$$\langle R(X, Y)Y, X \rangle = K_0 \cdot \|X, Y\|^2,$$

where $\|X, Y\|$ is the area of the parallelogram spanned by X and Y . Applying the formula on pg. I.308 (to the subspace of M_p spanned by X and Y), we can write this equation as

$$\langle R(X, Y)Y, X \rangle = K_0[\langle X, X \rangle \langle Y, Y \rangle - \langle X, Y \rangle^2].$$

18. LEMMA. If $p \in M$ is isotropic, with all sectional curvatures equal to K_0 , then for all $X, Y, Z, W \in M_p$ we have

$$\langle R(X, Y)Z, W \rangle = K_0[\langle X, W \rangle \langle Y, Z \rangle - \langle X, Z \rangle \langle Y, W \rangle].$$

PROOF. Denote the right side of this equation by $\mathcal{R}(X, Y, Z, W)$. By hypothesis, $\langle R(X, Y)Y, X \rangle = \mathcal{R}(X, Y, Y, X)$. It is easy to check that \mathcal{R} has properties (1)–(3), and hence (4), of Proposition 4-11. The desired result now follows from Proposition 4-12. ♦

Although not necessary for our calculations, the following consequence of this Lemma is a standard result, and is pertinent to the topic of constant curvature manifolds.

19. THEOREM (SCHUR). If M is a connected Riemannian manifold of dimension $n \geq 3$ and all points of M are isotropic, then M has constant curvature.

PROOF. By Lemma 18 we have, in a coordinate system x ,

$$R_{hijk} = K[g_{hj}g_{ik} - g_{hk}g_{ij}]$$

for some function K on M . Equivalently,

$$R^h_{ijk} = K(\delta_j^h g_{ik} - \delta_k^h g_{ij}).$$

Consequently, Ricci's Lemma (Proposition 5-3) and the relation $\delta_{j;l}^h = 0$ imply that

$$R^h_{ijk;l} = \frac{\partial K}{\partial x^l}(\delta_j^h g_{ik} - \delta_k^h g_{ij}).$$

From Bianchi's identity (Proposition 5-9(3')) we then obtain

$$0 = \frac{\partial K}{\partial x^l} (\delta_j^h g_{ik} - \delta_k^h g_{ij}) + \frac{\partial K}{\partial x^j} (\delta_k^h g_{il} - \delta_l^h g_{ik}) + \frac{\partial K}{\partial x^k} (\delta_l^h g_{ij} - \delta_j^h g_{il}).$$

In this identity set $h = k$ and sum over k , to obtain

$$\begin{aligned} 0 &= \frac{\partial K}{\partial x^l} (g_{ij} - n g_{ij}) + \frac{\partial K}{\partial x^j} (n g_{il} - g_{il}) + \frac{\partial K}{\partial x^l} g_{ij} - \frac{\partial K}{\partial x^j} g_{il} \\ &= (n - 2) \cdot \left[\frac{\partial K}{\partial x^j} g_{il} - \frac{\partial K}{\partial x^l} g_{ij} \right]. \end{aligned}$$

Since $n \geq 3$ we have

$$\frac{\partial K}{\partial x^j} g_{il} = \frac{\partial K}{\partial x^l} g_{ij}.$$

Hence

$$\delta_l^m \frac{\partial K}{\partial x^j} = \sum_i g^{mi} g_{il} \frac{\partial K}{\partial x^j} = \sum_i g^{mi} g_{ij} \frac{\partial K}{\partial x^l} = \delta_j^m \frac{\partial K}{\partial x^l}.$$

Choosing $m = j \neq l$, we obtain $\partial K / \partial x^l = 0$, for all l . So K is constant. ♦

To obtain the metric in a space of constant curvature K_0 , we now consider the equations of Corollary 9. By Lemma 18 we have

$$\mathbf{R}^i_{jkl} = \langle R(X_k, X_l)X_j, X_i \rangle = K_0[\delta_{ki}\delta_{lj} - \delta_{kj}\delta_{li}],$$

which implies that

$$\mathbf{R}^i_{jij} = -\mathbf{R}^i_{jji} = K_0 \quad (j \neq i); \quad \text{all other } \mathbf{R}^i_{jkl} = 0.$$

So our equations become

$$(*) \quad \frac{\partial^2 \bar{\theta}^i}{\partial t^2} = -K_0 \sum_k t^k (t^k \bar{\theta}^i - t^i \bar{\theta}^k).$$

To solve even these simplified equations requires quite a bit of trickiness. We first use the equations in Proposition 8 to obtain

$$\frac{\partial \left(\sum_i t^i \bar{\theta}^i \right)}{\partial t} = \sum_i t^i dt^i + \sum_{i,k} t^k t^i \bar{\omega}_k^i,$$

and thus

$$\frac{\partial \left(\sum_i t^i \bar{\theta}^i \right)}{\partial t} = \sum_i t^i dt^i, \quad \text{using skew-symmetry of the } \bar{\omega}_k^i = \Phi^* \omega_k^i.$$

Since $\sum_i t^i \bar{\theta}^i(0, X) = 0$, we have

$$\sum_i t^i \theta^i = t \sum_i t^i dt^i,$$

and hence

$$(**) \quad \left(\sum_i t^i \bar{\theta}^i \right)^2 = \sum_{i,j} t^i t^j \bar{\theta}^i \bar{\theta}^j = t^2 \sum_{i,j} t^i t^j dt^i dt^j = t^2 \left(\sum_i t^i dt^i \right)^2.$$

We next obtain equations for the quantities $t^i \bar{\theta}^j - t^j \bar{\theta}^i$. By (*) we have

$$\begin{aligned} \frac{\partial^2 (t^i \bar{\theta}^j - t^j \bar{\theta}^i)}{\partial t^2} &= -K_0 \sum_{k=1}^n t^k t^i (t^k \bar{\theta}^j - t^j \bar{\theta}^k) + t^k t^j (t^k \bar{\theta}^i - t^i \bar{\theta}^k) \\ &= -K_0 \sum_{k=1}^n (t^k)^2 (t^i \bar{\theta}^j - t^j \bar{\theta}^i) \\ &= -K_0 \rho^2 (t^i \bar{\theta}^j - t^j \bar{\theta}^i), \end{aligned}$$

where we have set

$$\rho = \sqrt{\sum_{k=1}^n (t^k)^2}.$$

Together with the initial conditions

$$\begin{aligned} (t^i \bar{\theta}^j - t^j \bar{\theta}^i)(0, X) &= 0 \\ \frac{\partial (t^i \bar{\theta}^j - t^j \bar{\theta}^i)}{\partial t}(0, X) &= t^i dt^j - t^j dt^i, \end{aligned}$$

we obtain

$$t^i \bar{\theta}^j - t^j \bar{\theta}^i = \frac{s(\rho \sqrt{|K_0|} t)}{\rho \sqrt{|K_0|}} (t^i dt^j - t^j dt^i).$$

Recall that for a form η , we use η^2 to denote the quadratic function $X \mapsto \eta(X) \cdot \eta(X)$. Summing the squares of the equations just derived, we now obtain

$$\begin{aligned}
& \frac{s^2(\rho\sqrt{|K_0|}t)}{\rho^2|K_0|} \sum_{i < j} (t^i dt^j - t^j dt^i)^2 \\
&= \frac{1}{2} \frac{s^2(\rho\sqrt{|K_0|}t)}{\rho^2|K_0|} \sum_{i,j} (t^i dt^j - t^j dt^i)^2 \\
&= \frac{1}{2} \sum_{i,j} (t^i \bar{\theta}^j - t^j \bar{\theta}^i)^2 = \frac{1}{2} \left[\sum_{i,j} (t^i)^2 (\bar{\theta}^j)^2 + (t^j)^2 (\bar{\theta}^i)^2 - 2t^i t^j \bar{\theta}^i \bar{\theta}^j \right] \\
&= \sum_{i,j} (t^i)^2 (\bar{\theta}^j)^2 - t^2 \sum_{i,j} t^i t^j dt^i dt^j \quad \text{by (**)} \\
&= \rho^2 \sum_{k=1}^n (\bar{\theta}^k)^2 - t^2 \left(\sum_k t^k dt^k \right)^2.
\end{aligned}$$

Since $\exp: (M_p, \sum_k \bar{\theta}^k \otimes \bar{\theta}^k) \rightarrow (M, \langle \cdot, \cdot \rangle)$ is an isometry, and $\bar{\theta}$ is the value of $\bar{\theta}^i$ when $t = 1$, the expression for $\| \cdot \|^2$ in normal coordinates is obtained by seeing what $\sum_k (\bar{\theta}^k)^2$ becomes when we set $t = 1$ and $t^k = x^k$ in the above equation. We let r denote what ρ becomes when we perform this substitution; that is, we let

$$r = \sqrt{\sum_{k=1}^n (x^k)^2}.$$

We then obtain

$$\| \cdot \|^2 = \frac{1}{r^2} \left\{ \left(\sum_k x^k dx^k \right)^2 + \frac{s^2(\sqrt{|K_0|}r)}{|K_0|r^2} \cdot \sum_{i < j} (x^i dx^j - x^j dx^i)^2 \right\}.$$

Notice that we have

$$\left(\sum_k x^k dx^k \right)^2 = r^2 \left[\sum_k (dx^k)^2 \right] - \sum_{i < j} (x^i dx^j - x^j dx^i)^2;$$

consequently, we can write

$$\| \|^2 = \sum_k (dx^k)^2 - \left[\frac{|K_0|r^2 - s^2(\sqrt{|K_0|}r)}{|K_0|r^4} \right] \sum_{i < j} (x^i dx^j - x^j dx^i)^2.$$

From the Taylor series for \sin it is easy to see that the coefficient of the sum $\sum_{i < j} (x^i dx^j - x^j dx^i)^2$ is C^∞ (in fact it is analytic). This formula thus gives a direct verification of Riemann's assertion about the form of $\| \|^2$ in normal coordinates (page 168). On the other hand, this form of the metric is *not* the one which Riemann mentioned in his lecture (page 159). We will consider this metric in Addendum 2.

ADDENDUM 2

CONFORMALLY EQUIVALENT MANIFOLDS

In a vector space V with a positive definite inner product $\langle \cdot, \cdot \rangle$ we define the **angle** $\angle(v, w)$ between two non-zero vectors $v, w \in V$ by

$$\angle(v, w) = \arccos \frac{\langle v, w \rangle}{\|v\| \cdot \|w\|}.$$

It is easy to say when two metrics give the same angle measurements.

20. LEMMA. Let $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ be two positive definite inner products on V . Then $\angle_1(v, w) = \angle_2(v, w)$ for all non-zero $v, w \in V$ if and only if there is a number $c > 0$ with $\langle \cdot, \cdot \rangle_2 = c \cdot \langle \cdot, \cdot \rangle_1$.

PROOF. If $\langle \cdot, \cdot \rangle_2 = c \cdot \langle \cdot, \cdot \rangle_1$, then clearly $\angle_2(v, w) = \angle_1(v, w)$ for all $v, w \neq 0$.

Conversely, suppose $\angle_2(v, w) = \angle_1(v, w)$ for all $v, w \neq 0$. Let v_1, \dots, v_n be an orthonormal basis for V with respect to $\langle \cdot, \cdot \rangle_1$. Then for $i \neq j$ we have

$$\angle_2(v_i, v_j) = \angle_1(v_i, v_j) = 0,$$

so $\langle v_i, v_j \rangle_2 = 0$. Define c_i by $\langle v_i, v_i \rangle_2 = c_i$. Then

$$\begin{aligned} \angle_1(v_i, v_i + v_j) &= \frac{\langle v_i, v_i + v_j \rangle_1}{\sqrt{\langle v_i, v_i \rangle_1} \sqrt{\langle v_i + v_j, v_i + v_j \rangle_1}} = \frac{1}{\sqrt{2}}, \\ \angle_2(v_i, v_i + v_j) &= \frac{\langle v_i, v_i + v_j \rangle_2}{\sqrt{\langle v_i, v_i \rangle_2} \sqrt{\langle v_i + v_j, v_i + v_j \rangle_2}} = \frac{c_i}{\sqrt{c_i} \sqrt{c_i + c_j}}. \end{aligned}$$

It follows that $c_i = c_j$. ♦

Now a diffeomorphism $f: (M_1, \langle \cdot, \cdot \rangle_1) \rightarrow (M_2, \langle \cdot, \cdot \rangle_2)$ between Riemannian manifolds is called **conformal** if each f_{*p} is angle preserving, and the two

Riemannian manifolds are then called **conformally equivalent**. By Lemma 20, a diffeomorphism f is conformal if and only if

$$\langle \cdot, \cdot \rangle_1 = \lambda f^* \langle \cdot, \cdot \rangle_2$$

for some positive function $\lambda: M_1 \rightarrow \mathbb{R}$. In particular, if a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ is conformal to \mathbb{R}^n with its usual Riemannian metric, then around each point $p \in M$ we can choose a coordinate system x such that $g_{ij} = \lambda \delta_{ij}$ for some positive function $\lambda: M \rightarrow \mathbb{R}$. Such a coordinate system x is called **isothermal**.

In 1822 Gauss showed that isothermal coordinates can be found at any point of an arbitrary surface. His proof depends on a trick that works only for analytic (C^ω) manifolds, and uses a little knowledge of complex function theory. It is presented in Addendum 1 to Chapter 9 of Volume IV, together with a (much more involved) proof that works in the C^∞ case. For $n > 2$ it is not true that every n -manifold is locally conformally equivalent to \mathbb{R}^n . But certain important n -dimensional Riemannian manifolds are. In fact, in his lecture, Riemann states that on a manifold of constant curvature K_0 there is a coordinate system x with

$$\| \cdot \| = \frac{\sqrt{\sum_{i=1}^n (dx^i)^2}}{1 + \frac{K_0}{4} \sum_i (x^i)^2},$$

or equivalently

$$\langle \cdot, \cdot \rangle = \sum_{i=1}^n \frac{dx^i \otimes dx^i}{\left[1 + \frac{K_0}{4} \sum_i (x^i)^2\right]^2}.$$

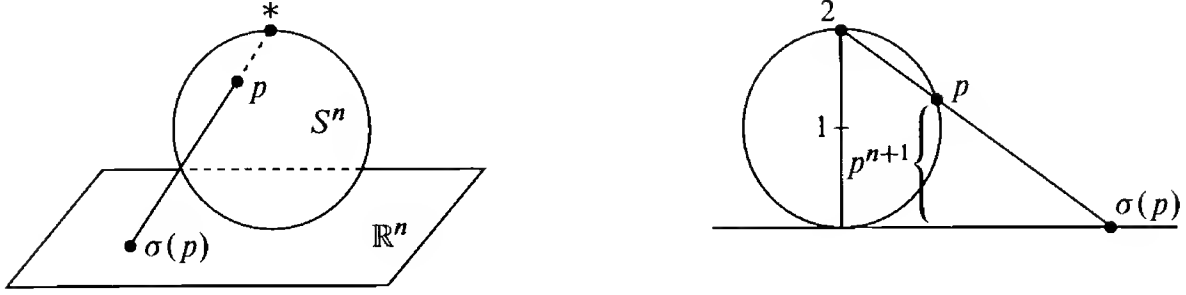
Thus, x is an isothermal coordinate system.

It wasn't very hard for Riemann to guess this, because there is a very standard coordinate system for S^n which gives this metric (with $K_0 = 1$). We consider S^n as the sphere of radius 1 around the point $(0, \dots, 0, 1)$, so that S^n is tangent to $\mathbb{R}^n = \mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+1}$. Let $*$ be the "north pole", $* = (0, \dots, 0, 2) \in S^n$. The **stereographic projection**

$$\sigma: S^n - \{*\} \rightarrow \mathbb{R}^n$$

is defined geometrically as follows: for any $p \neq *$ in S^n , we let $\sigma(p)$ be the

point where the line between p and $*$ intersects \mathbb{R}^n . It is easy to check (see the



plane section on the right of the diagram) that

$$(1) \quad \sigma(p) = \left(\frac{2p^1}{2 - p^{n+1}}, \dots, \frac{2p^n}{2 - p^{n+1}} \right)$$

and that $f = \sigma^{-1}$ is given by

$$(2) \quad \sigma^{-1}(y) = f(y) = \left(\frac{y^1}{1 + \frac{1}{4} \sum (y^i)^2}, \dots, \frac{y^n}{1 + \frac{1}{4} \sum (y^i)^2}, \frac{\frac{1}{2} \sum (y^i)^2}{1 + \frac{1}{4} \sum (y^i)^2} \right).$$

If y^1, \dots, y^n denotes the standard coordinate system on \mathbb{R}^n , then the $x^i = y^i \circ \sigma$ give a coordinate system on $S^n - \{*\}$.

Let $i: S^n \rightarrow \mathbb{R}^{n+1}$ be the inclusion map, and $\langle \cdot, \cdot \rangle = \sum_{i=1}^{n+1} dz^i \otimes dz^i$ the usual Riemannian metric on \mathbb{R}^{n+1} , so that $i^* \langle \cdot, \cdot \rangle$ is the usual Riemannian metric on S^n . We find the expression

$$i^* \langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j$$

for $i^* \langle \cdot, \cdot \rangle$ by writing

$$\begin{aligned} \sum_{i,j=1}^n (g_{ij} \circ f) dy^i \otimes dy^j &= f^* \left(\sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j \right) \\ &= f^* i^* \langle \cdot, \cdot \rangle = f^* \langle \cdot, \cdot \rangle \\ &= f^* \sum_{i=1}^{n+1} dz^i \otimes dz^i = \sum_{i=1}^{n+1} df^i \otimes df^i \\ &= \sum_{i=1}^{n+1} \sum_{j,k=1}^n \frac{\partial f^i}{\partial y^j} \frac{\partial f^i}{\partial y^k} dy^j \otimes dy^k, \end{aligned}$$

and computing the $\partial f^i / \partial y^j$ from (2). A forbidding looking computation finally coalesces into

$$\sum_{i,j=1}^n (g_{ij} \circ f) dy^i \otimes dy^j = \sum_{i=1}^n \frac{dy^i \otimes dy^i}{\left[1 + \frac{1}{4} \sum_i (y^i)^2\right]^2},$$

so that the expression for the metric $i^*\langle \ , \ \rangle$ is

$$\sum_{i=1}^n \frac{dx^i \otimes dx^i}{\left[1 + \frac{1}{4} \sum_i (x^i)^2\right]^2}.$$

(A neater way of deriving this formula is presented in Volume IV, Chapter 7.) If we were dealing with a sphere of curvature K_0 , the factor $1/4$ would be replaced by $K_0/4$.

Of course, we still have to check that this formula gives a metric of constant curvature K_0 when $K_0 < 0$. We will consider, more generally, all metrics on (a subset of) \mathbb{R}^n which have the standard coordinate system as isothermal coordinates, and determine just which of these have constant curvature.

Thus, we consider metrics on (a subset of) \mathbb{R}^n of the form

$$g_{ij} = \frac{\delta_{ij}}{F^2} \quad F \text{ nowhere } 0.$$

Then $g^{ij} = \delta^{ij} F^2$. We also have

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{-2\delta_{ij}}{F^3} \frac{\partial F}{\partial x^k} = \frac{-2\delta_{ij}}{F^2} \frac{\partial \log F}{\partial x^k}.$$

Setting $\log F = f$, and using the formulas on pp. I.326 and I.328, we obtain

$$\Gamma_{ii}^i = \frac{-\partial f}{\partial x^i}, \quad \Gamma_{jj}^j = \frac{\partial f}{\partial x^j}, \quad \Gamma_{ij}^i = \Gamma_{ji}^j = \frac{-\partial f}{\partial x^j} \quad (i \neq j); \quad \text{all other } \Gamma_{jk}^i = 0.$$

From the formula on page 214 we obtain

$$R^i_{jij} = -R^i_{jji} = \frac{\partial^2 f}{\partial x^j \partial x^j} + \frac{\partial^2 f}{\partial x^i \partial x^i} - \sum_{r \neq i,j} \left(\frac{\partial f}{\partial x^r} \right)^2 \quad i \neq j$$

$$R^i_{jjl} = \frac{-\partial^2 f}{\partial x^i \partial x^l} - \frac{\partial f}{\partial x^i} \frac{\partial f}{\partial x^l} \quad i, j, l \text{ distinct}$$

$$R^i_{jil} = \frac{\partial^2 f}{\partial x^j \partial x^l} + \frac{\partial f}{\partial x^j} \frac{\partial f}{\partial x^l} \quad i, j, l \text{ distinct}$$

$$\text{all other } R^i_{jkl} = 0.$$

Now by Lemma 18, the metric has constant curvature K_0 if and only if

$$R_{ijkl} = K_0(g_{ik}g_{jl} - g_{il}g_{jk}),$$

or equivalently

$$\begin{aligned} R^i_{jkl} &= K_0(\delta_k^i g_{jl} - \delta_l^i g_{jk}) \\ &= \frac{K_0}{F^2}(\delta_k^i \delta_{jl} - \delta_l^i \delta_{jk}), \end{aligned}$$

and hence

$$R^i_{jij} = -R^i_{jji} = \frac{K_0}{F^2} \quad (j \neq i); \quad \text{all other } R^i_{jkl} = 0.$$

So the metric has constant curvature K_0 if and only if

$$\begin{aligned} \frac{\partial^2 f}{\partial x^j \partial x^l} + \frac{\partial f}{\partial x^j} \frac{\partial f}{\partial x^l} &= 0 & j \neq l \\ \frac{\partial^2 f}{\partial x^i \partial x^i} + \frac{\partial^2 f}{\partial x^j \partial x^j} - \sum_{r \neq i, j} \left(\frac{\partial f}{\partial x^r} \right)^2 &= \frac{K_0}{F^2} & i \neq j. \end{aligned}$$

Since

$$\frac{\partial^2 f}{\partial x^j \partial x^l} + \frac{\partial f}{\partial x^j} \frac{\partial f}{\partial x^l} = \frac{1}{F} \frac{\partial^2 f}{\partial x^j \partial x^l} \quad \text{for all } j, l,$$

these equations hold if and only if

$$(1) \quad \frac{\partial^2 F}{\partial x^j \partial x^l} = 0 \quad j \neq l$$

$$(2) \quad F \cdot \left(\frac{\partial^2 F}{\partial x^i \partial x^i} + \frac{\partial^2 F}{\partial x^j \partial x^j} \right) = K_0 + \sum_{r=1}^n \left(\frac{\partial F}{\partial x^r} \right)^2 \quad i \neq j.$$

Equation (1) implies that $F = G_1 + \cdots + G_n$, where G_j depends only on x^j . Using equation (2) for i, l and then for j, l , we obtain

$$\frac{\partial^2 G_i}{\partial x^i \partial x^i} = \frac{\partial^2 G_j}{\partial x^j \partial x^j}.$$

So we must have

$$G_i = cx_i^2 + b_i x_i + a_i,$$

for some c . From (2) we then obtain

$$K_0 = \sum_{r=1}^n (4a_r c - b_r^2).$$

There are two important special cases. If we choose $c = K_0/4$, $b_i = 0$, $a_i = 1/n$ we obtain the metric $\langle \cdot, \cdot \rangle = \sum g_{ij} dx^i \otimes dx^j$ which Riemann mentions, with

$$g_{ij} = \frac{\delta_{ij}}{\left[1 + \frac{K_0}{4} \sum_i (x^i)^2\right]^2}.$$

Notice that for $K_0 < 0$, this metric is defined only on

$$M = \left\{ a \in \mathbb{R}^n : \sum_{i=1}^n (a^i)^2 < -4/K_0 \right\}.$$

Nevertheless, $(M, \langle \cdot, \cdot \rangle)$ is complete. To see this, we compute that the curve γ defined by

$$\begin{aligned} \gamma^1(t) &= 2 \frac{\sinh \frac{\sqrt{|K_0|} t}{2}}{\sqrt{|K_0|} \cosh \frac{\sqrt{|K_0|} t}{2}} \\ \gamma^i(t) &= 0, \quad i > 1 \end{aligned}$$

is a geodesic through 0, parameterized by arclength, and defined for all t . Since the metric $\langle \cdot, \cdot \rangle$ is radially symmetric around 0, there are geodesics through 0 in all directions, which are defined for all t .

When $K_0 > 0$, the same metric is defined on all of \mathbb{R}^n , but it is *not* complete, since it is isometric to the n -sphere of radius $\sqrt{K_0}$ with a point deleted. Some theorems from Chapters 7 and 8 of Volume IV will throw more light on these matters.

The other important case occurs when $c = a_i = 0$ and $b_i = \delta_{in}$, so that the metric is

$$\sum_{i=1}^n \frac{dx^i \otimes dx^i}{(x^n)^2},$$

with constant curvature $K_0 = -1$. The 2-dimensional case, in particular, gives the nicest example of a non-Euclidean geometry (see Problem I.9-41).

ADDENDUM 3

É. CARTAN'S TREATMENT
OF NORMAL COORDINATES

In this Addendum we use moving frames to prove Riemann's claims about the form of the metric in normal coordinates, and in fact obtain stronger results; a special case of our result is represented by the form of the metric for spaces of constant curvature which we found in Addendum 1. Throughout, we will be working with a moving frame X_1, \dots, X_n adapted to an orthonormal basis X_{1p}, \dots, X_{np} for M_p ; the forms $\theta^i, \bar{\theta}^i$ are as defined previously, and x^1, \dots, x^n denotes the Riemannian normal coordinate system determined by X_{1p}, \dots, X_{np} .

21. PROPOSITION. The quadratic form $\| \|^2 - \sum_i (dx^i)^2$ can be written as a quadratic form in the differentials $x^r dx^s - x^s dx^r$.

PROOF. Consider the functions A_{jk}^i on $\mathbb{R} \times M_p$ which satisfy the equations

$$(1) \quad \begin{cases} \frac{\partial^2 A_{jk}^i}{\partial t^2} = \sum_r (\mathbf{R}_{rjk}^i \circ \Phi) t t^r + \sum_{r,s,l} (\mathbf{R}_{rsl}^i \circ \Phi) A_{jk}^l t^r t^s \\ A_{jk}^i(0, X) = 0 \\ \frac{\partial A_{jk}^i}{\partial t}(0, X) = 0. \end{cases}$$

We claim that

$$(2) \quad \bar{\theta}^i = t dt^i + \sum_{j,k} A_{jk}^i t^j dt^k.$$

To prove this, we simply note that if we *define* $\bar{\theta}^i$ by this formula, then an easy calculation shows that the $\bar{\theta}^i$ satisfy the equations and initial conditions of Corollary 9; since the solutions are unique, the result follows.

Now note, by skew-symmetry of \mathbf{R}_{jkl}^i in the last two indices, that $A_{jk}^i + A_{kj}^i$ satisfies a *linear* second order differential equation, with initial conditions

$$\begin{aligned} (A_{jk}^i + A_{kj}^i)(0, X) &= 0 \\ \frac{\partial}{\partial t}(A_{jk}^i + A_{kj}^i)(0, X) &= 0. \end{aligned}$$

It follows that

$$(3) \quad A_{jk}^i = -A_{kj}^i.$$

[The motivation for this proof is the following. The equations in Proposition 8 clearly imply that $\bar{\theta}^i - t \, dt^i = 0$ at $(t, 0)$. This means (by Lemma I.3-2) that we can write $\bar{\theta}^i$ as in (2). Then the equations in Corollary 9 imply that the A_{jk}^i satisfy (1).]

Next consider the functions B_{sjk}^r which satisfy the equations

$$(4) \quad \left\{ \begin{array}{l} \frac{\partial^2 B_{sjk}^r}{\partial t^2} = (\mathbf{R}_{sjk}^r \circ \Phi)t + \sum_{\mu, l} \mathbf{R}_{s\mu l}^i A_{jk}^l t^\mu \\ B_{sjk}^r(0, X) = 0 \\ \frac{\partial B_{sjk}^r}{\partial t}(0, X) = 0. \end{array} \right.$$

We claim that

$$(5) \quad A_{jk}^r = \sum_s B_{sjk}^r t^s.$$

This is proved by checking that if we define A_{jk}^r by (5), then equations (4) imply equations (1).

Using the skew-symmetry of \mathbf{R}_{sjk}^r in j and k and the skew-symmetry of A_{jk}^l in j and k [equation (3)], we easily see that

$$(6) \quad B_{sjk}^r = -B_{skj}^r.$$

Also, using skew-symmetry of \mathbf{R}_{sjk}^r in r and s , we see that

$$(7) \quad B_{sjk}^r = -B_{rjk}^s.$$

We know that the expression for $\| \|^2 = \sum_i (\theta^i)^2$ in the Riemannian normal coordinate system x^1, \dots, x^n is what $\sum_i (\bar{\theta}^i)^2$ becomes when we set $t = 1$ and $t^i = x^i$. So by (2) we have

$$\| \|^2 = \sum_i (dx^i)^2 + \sum_i \left(\sum_{j, k} A_{jk}^i x^j dx^k \right)^2 + \sum_r \left(\sum_{j, k} A_{jk}^r dx^r x^j dx^k \right).$$

Now (3) implies that the first triple sum is a quadratic form in the $x^r dx^s - x^s dx^r$. The second triple sum can be written

$$\sum_{r,s,l,j} B'_{sjk} x^s dx^r x^j dx^k;$$

using (6) and (7), we see that this can also be written as a quadratic form in the $x^r dx^s - x^s dx^r$. ♦

For further developments, see É. Cartan's *Leçons sur la Géométrie des Espaces de Riemann*, pp. 242ff.

CHAPTER 8

CONNECTIONS IN PRINCIPAL BUNDLES

The method of moving frames turns out to be surprisingly powerful, but it leads us to a definition of a connection which does not have the “invariance” property of the Koszul definition (although it is a big improvement over the classical definition, simply because the transformation law can be stated so much more elegantly). At the same time, we should note a certain deficiency in our proofs of the Test Case. We certainly have enough of them (six so far), and they use the integrability conditions $R = 0$ in many different ways. In the first proof we use the classical integrability theorem, and in the second and third proofs we essentially reprove this theorem. In the fourth proof we use the differential form version of the Frobenius Theorem, and in the fifth and sixth proofs we use the proof of the integrability theorem outlined in Problem I.6-8. However, in all this time we have never used the distribution formulation of the Frobenius Integrability Theorem (I.6-5), although this is the most geometric version of all.

These two phenomena will eventually turn out to be closely related, but for the present we will concentrate on the first problem. The main step in the solution of this problem was accomplished by Ehresmann* in 1950. With the advantages of hindsight we can reconstruct the solution in a way that makes it seem natural and almost obvious.

It will be helpful to begin by reconsidering the classical and modern definitions of vector fields. As we pointed out long ago, the snazziest modern definition of tangent vectors is essentially the same as the classical definition, as n -tuples of numbers which “transform” according to certain rules. On the other hand, the modern definition of a vector field represents a definite improvement over the classical definition. Instead of dealing with n -tuples of functions transforming according to certain rules, we make the set of tangent vectors into a new manifold, the tangent bundle, and define vector fields to be sections of this vector bundle. The idea behind the modern treatment of connections is to ob-

* Ehresmann, *Les connexions infinitesimales dans un espace fibre differentiable*, Colloque de Topologie, Bruxelles (1950), 29–55.

tain a bundle whose sections are just the moving frames on M . To do this, we will imitate the construction of the tangent bundle in a rather straightforward way.

Recall that a **frame** u for M_p is just an ordered basis $u = (u_1, \dots, u_n)$ for M_p . Let $F(M)$ denote the set of all frames u for all tangent spaces M_p . We call $F(M)$ the **bundle of frames** of M , and define $\pi: F(M) \rightarrow M$ to be the map which takes a basis u for M_p to $\pi(u) = p$. If (x, U) is a coordinate system on M and $p \in U$, then every frame $u = (u_1, \dots, u_n)$ for M_p can be written uniquely as

$$u_j = \sum_{i=1}^n x_j^i(u) \frac{\partial}{\partial x^i} \Big|_{\pi(u)}.$$

The matrix $(x_j^i(u))$ is non-singular, and any non-singular matrix can occur, so the map

$$u \mapsto (x^i(\pi(u)), x_j^i(u)) \in \mathbb{R}^n \times \text{GL}(n, \mathbb{R})$$

is a one-one map $x_\#$ from $\pi^{-1}(U)$ onto $x(U) \times \text{GL}(n, \mathbb{R})$. It is easy to see that if (y, V) is another coordinate system, then $y_\# \circ (x_\#)^{-1}$ is C^∞ , from $x_\#(\pi^{-1}(U \cap V))$ to $y_\#(\pi^{-1}(U \cap V))$. This means that we can make $F(M)$ into a C^∞ manifold in such a way that each $x_\#$ is a diffeomorphism; with this C^∞ structure on $F(M)$, the map π is clearly C^∞ . Notice, finally, that we have a C^∞ map $F(M) \times \text{GL}(n, \mathbb{R}) \rightarrow F(M)$, given by $(u, A) \mapsto u \cdot A$, where $(u \cdot A)_i = \sum_j A_i^j u_j$; we have $u \cdot A = u$ only when $A = I$, and the set of all $u \cdot A$ for $A \in \text{GL}(n, \mathbb{R})$ is just $\pi^{-1}(\pi(u))$.

Although we have called $F(M)$ the “bundle of frames”, it is clearly not a vector bundle; each fibre $\pi^{-1}(p)$ is diffeomorphic to $\text{GL}(n, \mathbb{R})$, rather than to some \mathbb{R}^N . However, $F(M)$ is another special sort of “bundle” which we will now define.

Let M be a C^∞ manifold, and G a Lie group. A (C^∞) **principal bundle over M , with group G** , is a triple (P, π, \cdot) where

- (1) P is a C^∞ manifold (the **total space** of the principal bundle)
- (2) $\pi: P \rightarrow M$ is a C^∞ map (the **projection map** of the bundle) onto M (the **base space** of the principal bundle), satisfying

$$\pi(u \cdot a) = \pi(u) \quad \text{for all } u \in P \text{ and } a \in G$$

- (3) the map \cdot (the **action** of G) is a C^∞ map $(u, a) \mapsto u \cdot a$ from $P \times G$ to P with

$$u \cdot (ab) = (u \cdot a) \cdot b \quad \text{for all } u \in P \text{ and } a, b \in G$$

such that the following “local triviality” condition holds:

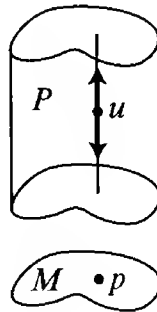
For each $p \in M$ there is a neighborhood U of p and a diffeomorphism $t: \pi^{-1}(U) \rightarrow U \times G$ of the form

$$t(u) = (\pi(u), \phi(u))$$

where ϕ satisfies $\phi(u \cdot a) = \phi(u)a$ [the latter product being the product in G].

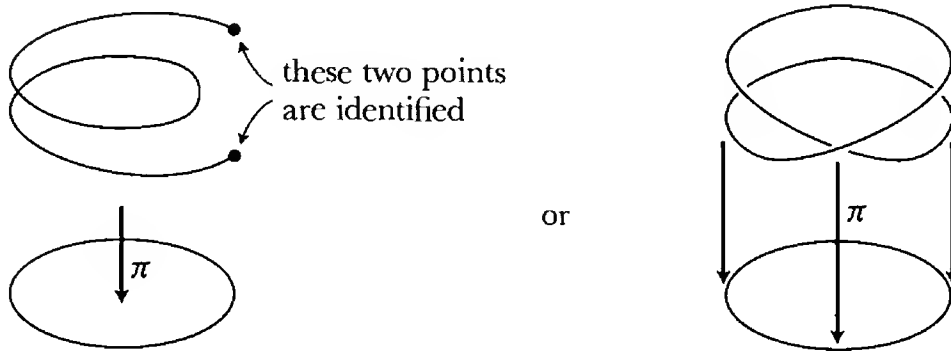
From the condition $\pi(u \cdot a) = \pi(a)$ we see that $\{u \cdot a : a \in G\} \subset \pi^{-1}(\pi(u))$. Using the property $\phi(u \cdot a) = \phi(u)a$ of the map ϕ , we see that we actually have $\{u \cdot a : a \in G\} = \pi^{-1}(\pi(u))$: for, if $v \in P$ satisfies $\pi(v) = \pi(u)$, and $\phi(v) = \phi(u)a$ for $a \in G$, then $\phi(v) = \phi(u \cdot a)$ and $\pi(v) = \pi(u \cdot a)$, so $v = u \cdot a$. Notice also that if $u \cdot a = u$ for some $u \in P$, then $a = e$.

Each “fibre” $\pi^{-1}(p)$ of P is clearly diffeomorphic to G . If $p = \pi(u)$ for $u \in P$, and $i: \pi^{-1}(p) \rightarrow P$ is the inclusion, then the image of the tangent space $i_*(\pi^{-1}(p)_u)$ is a subspace V_u of P_u , called the **vertical subspace** at u ; tangent vectors in this subspace are called **vertical** tangent vectors at u . Clearly $Y \in V_u$ is vertical if and only if $\pi_* Y = 0$.



The simplest example of a principal bundle is $M \times G$, with $\pi: M \times G \rightarrow M$ the projection on the first factor, and $(p, a) \cdot b = (p, ab)$. This is called the **trivial** principal bundle with group G . So far, we have given only one other example of a principal bundle, the bundle of frames $F(M)$, with group $\text{GL}(n, \mathbb{R})$. However, we can use the construction of this bundle to acquire many other examples. If $\pi: E \rightarrow M$ is any C^∞ vector bundle over M , we can let $F(E)$ be the collection of all frames u for the vector space $\pi^{-1}(p)$, for all $p \in M$; the projection map $\varpi: F(E) \rightarrow M$ takes a frame u for $\pi^{-1}(p)$ into $\varpi(u) = p$. Consider, in particular, the Möbius strip as a 1-dimensional vector bundle $\pi: E \rightarrow S^1$ over S^1 . A frame in a 1-dimensional vector space is just a non-zero vector, so $F(E)$ consists of the Möbius strip with the 0-section deleted. This space is connected (cut a paper Möbius strip along the center if you don't believe it); more generally, a vector bundle $\pi: E \rightarrow M$ over a connected space M is orientable if and only if $F(E)$ is disconnected.

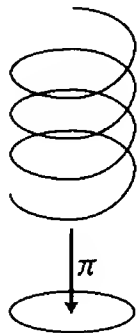
Examples of a different sort are obtained if we begin with a vector bundle $\pi : E \rightarrow M$ equipped with a Riemannian metric, and define $O(E)$ to be the set of all *orthonormal* frames u for $\pi^{-1}(p)$, for all $p \in M$. This is a principal bundle over M with group $O(n)$. In the case of a 1-dimensional bundle $\pi : E \rightarrow M$, every fibre of $O(E)$ has exactly 2 points, and $O(1) = \mathbb{Z}_2$; the action of the non-zero element of $O(1)$ on $O(E)$ interchanges these 2 points in each fibre. For the Möbius strip, the principal bundle $O(E)$ looks like the picture below.



More precisely, the total space of $O(E)$ is a circle S^1 , and the projection map $S^1 \rightarrow S^1$ is given by $\theta \mapsto 2\theta$.

In the case of an *oriented* bundle $\pi : E \rightarrow M$, these constructions can be modified to give principal bundles $SF(E)$ and $SO(E)$, the bundle of positively oriented frames, and positively oriented orthonormal frames, respectively. For 1-dimensional bundles, the total space $SF(E)$ looks like $M \times (0, \infty)$, while $SO(E)$ looks just like M . For 2-dimensional bundles, $SO(E)$ has a circle above each point of M ; the action of $\theta \in SO(2) = S^1$ rotates each circle through an angle of θ .

Although examples in which the group G is discrete are the least interesting for us, we will nevertheless give two more, as examples of principal bundles are hard to come by. In the first example, the group is the integers \mathbb{Z} , the total space and base space are \mathbb{R} and S^1 , respectively, and the map $\pi : \mathbb{R} \rightarrow S^1$ is



$\pi(\theta) = (\cos \theta, \sin \theta)$. In the second example, we define $\pi : S^2 \rightarrow \mathbb{P}^2$ to be

$\pi(p) = \{p, -p\}$; the group is \mathbb{Z}_2 and the action of the non-zero element is to take $p \in S^2$ to $-p$.

Before discussing principal bundles in particular, we need some generalities about Lie groups acting on manifolds. Consider a Lie group G , a C^∞ manifold M , and a C^∞ map $(p, a) \mapsto p \cdot a$ from $M \times G$ to M . We say that G acts on M on the right (via this map) if

- (1) the map $R_a: M \rightarrow M$ defined by $R_a(p) = p \cdot a$ is a diffeomorphism for all $a \in G$
- (2) $p \cdot (ab) = (p \cdot a) \cdot b$ for all $p \in M$ and $a, b \in G$.

Condition (2) can also be written as $R_{ab} = R_b \circ R_a$; since each R_a is a diffeomorphism, it follows easily that R_e is the identity map of M . We say that G acts **effectively** if e is the only element a with R_a the identity map of M , and we say that G acts **without fixed point** if the following stronger condition holds: if $p \cdot a = p$ for some $p \in M$, then $a = e$.

Now let \mathfrak{g} be the Lie algebra of a Lie group G which acts on M on the right. For every $X \in \mathfrak{g}$, we have the curve $t \mapsto \exp tX$ in G ; for each $p \in M$ this gives rise to a curve $c_p(t) = p \cdot (\exp tX) = R_{\exp tX}(p)$. We denote $c_p'(0)$ by $\sigma(X)(p)$; we thus have a vector field $\sigma(X)$ on M , and hence a map $\sigma: \mathfrak{g} \rightarrow$ (vector fields on M). The 1-parameter group of diffeomorphisms generated by $\sigma(X)$ is $\phi_t(p) = p \cdot (\exp tX)$, by the very definition of $\sigma(X)$. It is important to note that we can also describe $\sigma(X)$ as follows. For $p \in M$, let $\sigma_p: G \rightarrow M$ be $\sigma_p(a) = p \cdot a$. Then

$$\sigma(X)(p) = \sigma_{p*}(X).$$

To discuss this operation σ we will also need to introduce an important map, the “adjoint map”

$$\text{Ad}(a) = (L_a R_a^{-1})_* = (R_a^{-1} L_a)_*: \mathfrak{g} \rightarrow \mathfrak{g},$$

where L_a and R_a now denote left and right translations in G . Thus $\text{Ad}(a)$ is the differential at e of the map $b \mapsto aba^{-1} = L_a R_a^{-1}(b) = R_a^{-1} L_a(b)$. Usually $(\text{Ad}(a))(X)$ is denoted simply by $\text{Ad}(a)X$. If \tilde{X} is the left invariant vector field on G with $\tilde{X}(e) = X \in \mathfrak{g}$, then

$$\text{Ad}(a)X = (R_a^{-1})_*(L_{a*}\tilde{X})(e) = [(R_a^{-1})_*\tilde{X}](e),$$

since $L_{a*}\tilde{X} = \tilde{X}$. Consider, in particular, the special case where $G = \text{GL}(n, \mathbb{R})$, so that $\mathfrak{g} = \mathfrak{gl}(n, \mathbb{R})$ is the set of all $n \times n$ matrices. For any $n \times n$ matrix N and any $A \in \text{GL}(n, \mathbb{R})$ we have

$$\text{Ad}(A)N = ANA^{-1},$$

since $L_{A*} = L_A$ and $R_{A*} = R_A$, because L_A and R_A are linear functions (compare Problem I.10-19).

1. PROPOSITION. Let G act on the right on M . Then

- (1) The map $\sigma: \mathfrak{g} \rightarrow (\text{vector fields on } M)$ is linear.
- (2) $\sigma([X, Y]) = [\sigma(X), \sigma(Y)]$.
- (3) If G acts effectively and $X \neq 0$, then $\sigma(X)$ is not the zero vector field.
- (4) If G acts without fixed point and $X \neq 0$, then $\sigma(X)$ is nowhere 0.

PROOF. Linearity is clear from the equation $\sigma(X)(p) = \sigma_{p*}(X)$.

To prove (2) we note that since the bracket of two vector fields is the same as the Lie derivative (Theorem I.5-10), we have

$$\begin{aligned}
 (1) \quad [X, Y] &= [\tilde{X}, \tilde{Y}](e) = \lim_{h \rightarrow 0} \frac{1}{h} [Y - (R_{\exp hX})_* \tilde{Y}(e)] \\
 &= \lim_{h \rightarrow 0} \frac{1}{h} [Y - \text{Ad}(\exp -hX)Y].
 \end{aligned}$$

(Compare Problem I.10-19.)

On the other hand, if R_a now denotes the map $p \mapsto p \cdot a$ from M to G , then

$$(2) \quad R_{\exp hX} \circ \sigma_{p \cdot (\exp -hX)}(a) = p \cdot ([\exp -hX]a \exp hX).$$

Since $\phi_t(p) = p \cdot (\exp tX)$ is the 1-parameter group of diffeomorphisms generated by $\sigma(X)$, on M we have

$$\begin{aligned}
 [\sigma(X), \sigma(Y)](p) &= \lim_{h \rightarrow 0} \frac{1}{h} [\sigma(Y)(p) - [R_{\exp hX}]_* \sigma(Y)(p)] \\
 &= \lim_{h \rightarrow 0} \frac{1}{h} [\sigma_{p*} Y - \sigma_p(\text{Ad}(\exp -hX)Y)] && \text{by (2)} \\
 &= \sigma_{p*} \left(\lim_{h \rightarrow 0} \frac{1}{h} [Y - \text{Ad}(\exp -hX)Y] \right) = \sigma_{p*}([X, Y]) && \text{by (1)} \\
 &= \sigma([X, Y]).
 \end{aligned}$$

To prove (3), suppose $\sigma(X) = 0$. Then for every p the 1-parameter group of diffeomorphisms $\phi_t(p) = p \cdot (\exp tX)$ must be $\phi_t(p) = p$ (for the unique integral curve c of $\sigma(X)$ through p is clearly $c(p) = p$). If G acts effectively, this means that all $\exp tX = 0$, so $X = 0$.

To prove (4), suppose $\sigma(X)(p) = 0$ for some p . Then $p \cdot (\exp tX) = p$ for all t . If G acts without fixed point, then all $\exp tX = e$, so $X = 0$. ♦

Let us now apply this construction to a principal bundle $\pi: P \rightarrow M$, with group G . The map \cdot from $P \times G \rightarrow P$ is an action of G on P to the right (the map $u \mapsto u \cdot a$ is a diffeomorphism by condition (3) of the definition) and G acts without fixed point (we have already pointed out that this follows from condition (3)). Therefore we have the **fundamental vector field** $\sigma(X)$ corresponding to X for all $X \in \mathfrak{g}$; for every $u \in P$, the map $X \mapsto \sigma(X)(u)$ is an isomorphism, since G acts without fixed point. Since the maps $R_a: P \rightarrow P$ take fibres to themselves, the set of all $\sigma(X)(u)$ is precisely the set of vertical vectors at u .

2. PROPOSITION. For all $X \in \mathfrak{g}$ and $a \in G$, the vector field $(R_a)_*\sigma(X)$ is the fundamental vector field

$$(R_a)_*\sigma(X) = \sigma(\text{Ad}(a^{-1})X).$$

PROOF. Since $\phi_t(p) = p \cdot \exp tX = R_{\exp tX}(p)$ is the 1-parameter group of diffeomorphisms generated by $\sigma(X)$, it follows from Lemma I.5-11 that $(R_a)_*\sigma(X)$ generates the 1-parameter group of diffeomorphisms

$$\psi_t = R_a \circ R_{\exp tX} \circ R_a^{-1} = R_{a^{-1} \exp tX a}.$$

Now $\{a^{-1} \exp tX a\}$ is the 1-parameter group of diffeomorphisms of G generated by $\text{Ad}(a^{-1})X$. So ψ_t generates $\sigma(\text{Ad}(a^{-1})X)$. ♦

The vector fields $\sigma(X)$ are rather difficult to picture, especially since most principal bundles themselves are impossible to visualize. Nevertheless they will be very important, as we shall see upon returning to the structure which began our whole discussion, the principal bundle of frames $F(M)$. A **section** s of this bundle over an open set $U \subset M$ is a C^∞ map $s: U \rightarrow F(M)$ with $\pi \circ s = \text{identity map of } U$. Clearly, a section s is just what we used to call a moving frame on U . Note that, unlike a vector bundle, which always has a section defined on all of M , namely the 0-section, a principal bundle need not have such a section. In fact, if the principal bundle $\pi: P \rightarrow M$ over M with group G has a section $s: M \rightarrow P$, then the bundle is trivial: we can let $(p, a) \in M \times G$ correspond to $s(p) \cdot a \in P$.

We can use some of this new language to give an alternative, but completely equivalent, definition of a Cartan connection: A Cartan connection is an assignment of an $n \times n$ matrix-valued 1-form $\omega_s [= (\omega_{sj}^i)]$ to every section $s: U \rightarrow F(M)$ in such a way that

$$(*) \quad \omega_{s \cdot a} = a^{-1} da + a^{-1} \omega_s a$$

for every C^∞ function $a: U \rightarrow \text{GL}(n, \mathbb{R})$. (Here $s \cdot a$ is the section $(s \cdot a)(p) = s(p) \cdot a(p)$, the \cdot denoting the action of $\text{GL}(n, \mathbb{R})$ on $F(M)$.) This formulation of the definition suggests how we may obtain a definition of a connection which has all the advantages of the Cartan definition but which is also “invariant”. We ask if there is an $n \times n$ matrix-valued 1-form ω on the manifold $F(M)$ such that for each section (moving frame) s we have

$$\omega_s = s^*(\omega);$$

notice that previously we used ω alone to denote the connection form for some moving frame, but from now on we will have to be careful to use subscripts to distinguish the forms on M from the form ω which we hope to find on $F(M)$. Since the forms ω_s for a Cartan connection satisfy $(*)$, our question is then the following:

If we are given a collection of ω_s satisfying $(*)$, is there an ω on $F(M)$ such that each $\omega_s = s^*(\omega)$? More generally, which $n \times n$ matrix-valued 1-forms ω on $F(M)$ have the property that for every section $s: U \rightarrow F(M)$ and $\text{GL}(n, \mathbb{R})$ -valued function a on U we have

$$(**) \quad (s \cdot a)^*(\omega) = a^{-1} da + a^{-1} s^*(\omega) \cdot a?$$

In order to answer these questions, we need to know something about the section $s \cdot a$. If a has the constant value $A \in \text{GL}(n, \mathbb{R})$, then $s \cdot a = R_A \circ s$, where $R_A(u) = u \cdot A$, the dot denoting the action of $\text{GL}(n, \mathbb{R})$ on $F(M)$. So for any tangent vector $X_p \in M_p$ we have $(s \cdot a)_*(X_p) = R_{A*}(s_* X_p)$; when a is not constant there is a correction term.

3. PROPOSITION. Let s be a section of $F(M)$, over some open set U , and let $a: U \rightarrow \text{GL}(n, \mathbb{R})$ be C^∞ . Then for any tangent vector X_p at a point $p \in U$ we have

$$(s \cdot a)_*(X_p) = R_{a(p)*}(s_* X_p) + \sigma(a(p)^{-1} \cdot X_p(a))(s(p) \cdot a(p)).$$

[Note that $X_p(a)$ is an $n \times n$ matrix, so that $a(p)^{-1} \cdot X_p(a)$ is also an $n \times n$ matrix, and hence may be considered as an element of $\mathfrak{gl}(n, \mathbb{R})$, so $\sigma(a(p)^{-1} \cdot X_p(a))$ is a vector field on $F(M)$.]

PROOF. For convenience, let

$$m: F(M) \times \text{GL}(n, \mathbb{R}) \rightarrow F(M)$$

be $m(u, A) = u \cdot A$. Remember that the tangent space W of $F(M) \times \text{GL}(n, \mathbb{R})$ at (u, A) is isomorphic to the direct sum $F(M)_u \oplus \text{GL}(n, \mathbb{R})_A$, so we can consider

every element of W as a pair $(Y_1, Y_2) = Y_1 \oplus Y_2$, where $Y_1 \in F(M)_u$ and $Y_2 \in \text{GL}(n, \mathbb{R})_A$. Note that if

$$\begin{aligned} c_1 & \text{ is an integral curve in } F(M) && \text{for } Y_1 \\ c_2 & \text{ is an integral curve in } \text{GL}(n, \mathbb{R}) && \text{for } Y_2, \end{aligned}$$

then

$$\begin{aligned} t \mapsto m(c_1(t), A) = c_1(t) \cdot A & \text{ is an integral curve for } X \oplus 0 \\ t \mapsto m(u, c_2(t)) = u \cdot c_2(t) & \text{ is an integral curve for } 0 \oplus Y. \end{aligned}$$

Now let c be an integral curve for X_p . Since $s \cdot a = m \circ (s, a)$, we have

$$\begin{aligned} (s \cdot a)_*(X_p) &= m_*(s_*(X_p), a_*(X_p)) = m_*(s_*(X_p) \oplus a_*(X_p)) \\ &= \left. \frac{d}{dt} \right|_{t=0} s(c(t)) \cdot a(p) + \left. \frac{d}{dt} \right|_{t=0} s(p) \cdot a(c(t)). \end{aligned}$$

The first term on the right can be written as

$$\begin{aligned} \left. \frac{d}{dt} \right|_{t=0} R_{a(p)}(s(c(t))) &= R_{a(p)*} \left(\left. \frac{ds(c(t))}{dt} \right|_{t=0} \right) \\ &= R_{a(p)*}(s_*X_p). \end{aligned}$$

To identify the second term, recall that for all $u \in F(M)$ we have

$$\sigma(N)(u) = \sigma_{u*}(N), \quad \text{where } \sigma_u(A) = u \cdot A \text{ for } A \in \text{GL}(n, \mathbb{R}).$$

We write

$$s(p) \cdot a(c(t)) = s(p) \cdot a(p) \cdot [a(p)^{-1} \cdot a(c(t))].$$

The term in brackets gives a curve $\gamma(t) = a(p)^{-1} \cdot a(c(t))$ in $\text{GL}(n, \mathbb{R})$ with $\gamma(0) = I$; then $\gamma'(0) \in \text{GL}(n, \mathbb{R})_I = \mathfrak{gl}(n, \mathbb{R})$ is the tangent vector corresponding to the matrix

$$\left. \frac{d}{dt} \right|_{t=0} a(p)^{-1} \cdot a(c(t)) = a(p)^{-1} \cdot \left. \frac{d}{dt} \right|_{t=0} a(c(t)) = a(p)^{-1} \cdot X_p(a),$$

when we identify the $n \times n$ matrices with $\mathfrak{gl}(n, \mathbb{R})$. Consequently,

$$\begin{aligned} \left. \frac{d}{dt} \right|_{t=0} s(p) \cdot a(c(t)) &= \left. \frac{d}{dt} \right|_{t=0} s(p) \cdot a(p) \cdot [a(p)^{-1} a(c(t))] \\ &= \sigma_{s(p) \cdot a(p)*} \left(\left. \frac{d}{dt} \right|_{t=0} a(p)^{-1} a(c(t)) \right) \\ &= \sigma_{s(p) \cdot a(p)*}(a(p)^{-1} \cdot X_p(a)) \\ &= \sigma(a(p)^{-1} \cdot X_p(a))(s(p) \cdot a(p)). \quad \spadesuit \end{aligned}$$

Remark: Proposition 3 can actually be formulated for any principal bundle $\pi: P \rightarrow M$ over M , with group G . If s is a section over U and $a: U \rightarrow G$ is C^∞ , then for any tangent vector X_p at $p \in U$ we have

$$(s \cdot a)_*(X_p) = R_{a(p)*}(s_*X_p) + \sigma(L_{a(p)^{-1}*}a_*(X_p))(s(p) \cdot a(p)).$$

With Proposition 3 at hand, let us reconsider our question. We want to know which $n \times n$ matrix-valued 1-forms ω on $F(M)$ have the property that for each section s we have

$$(**) \quad (s \cdot a)^*(\omega) = a^{-1}da + a^{-1}s^*(\omega)a.$$

This is equivalent to saying that for every $X_p \in M_p$ we have

$$\omega((s \cdot a)_*X_p) = a^{-1}(p)X_p(a) + a^{-1}(p)\omega(s_*X_p)a(p).$$

According to Proposition 3 this is equivalent to

$$\begin{aligned} \omega(R_{a(p)*}(s_*X_p)) + \omega(\sigma(a(p)^{-1} \cdot X_p(a))(s(p) \cdot a(p))) \\ = a^{-1}(p)X_p(a) + a^{-1}(p)\omega(s_*X_p)a(p). \end{aligned}$$

We can extract two separate equations from this, as follows:

(I) First suppose that $a(p) = I$. Then we obtain

$$\omega(s_*X_p) + \omega(\sigma(X_p(a))(s(p))) = X_p(a) + \omega(s_*X_p),$$

or

$$(I) \quad \omega(\sigma(X_p(a))(s(p))) = X_p(a) \quad \text{for } a(p) = I.$$

(2) Now suppose a has the constant value A . Then we obtain

$$(II) \quad \omega(R_{A*}(s_*X_p)) = A^{-1}\omega(s_*X_p)A.$$

Conversely, it is not hard to show that if ω satisfies (I) and (II), then ω satisfies (**).

Now the equations (I) and (II) can be simplified considerably, resulting in equations that do not involve s at all. In (I), the matrix $X_p(a)$ can obviously be any $n \times n$ matrix, since we just have to satisfy $a(p) = I$. Since we can also choose any s , we see that (I) is equivalent to

$$(I') \quad \omega(\sigma(N)(u)) = N \quad \text{for all } N \in \mathfrak{gl}(n, \mathbb{R}) \text{ and } u \in F(M).$$

We claim that (II) is equivalent to

$$(II') \quad \omega(R_{A*}Y) = A^{-1}\omega(Y)A \quad \text{for all } Y \in F(M)_u.$$

To see this we note that by choosing s appropriately we can make s_*X_p be any vector in $u = s(p)$ that is not vertical; since the vertical vectors are all of the form $\sigma(N)$, equation (II') for vertical vectors follows from (I') and Proposition 2, remembering that for $GL(n, \mathbb{R})$ we have $\text{Ad}(A)N = ANA^{-1}$.

Summing up, we see that a matrix-valued 1-form ω on $F(M)$ satisfies

$$(s \cdot a)^*(\omega) = a^{-1}da + a^{-1}s^*(\omega)a$$

[and consequently the assignment of $s^*\omega$ to s is a Cartan connection] if and only if

$$\begin{aligned} \omega(\sigma(N)) &= N & \text{for all } N \in \mathfrak{gl}(n, \mathbb{R}) \\ \omega(R_{A*}(Y)) &= A^{-1}\omega(Y)A = \text{Ad}(A^{-1})\omega(Y) & \text{for all } Y \in F(M)_u. \end{aligned}$$

We leave it to the reader to show that if we are given a Cartan connection $\{\omega_s\}$, then there is a unique such ω on $F(M)$ with $\omega_s = s^*(\omega)$. (Define $\omega(Y_u) = \omega_s(X_p)$ whenever $Y_u = s_*X_p$, and use Proposition 3 and the transformation rules for a Cartan connection to verify that ω is well-defined.) We are consequently ready for the final definition of a connection; since all our conditions make sense in *any* principal bundle, our new definition is not only more abstract, more elegant, and more incomprehensible, but also more general.

An **(Ehresmann) connection** in a principal bundle $\pi: P \rightarrow M$ over M with group G is a C^∞ \mathfrak{g} -valued 1-form ω on P such that

- (1) $\omega(\sigma(X)) = X$ for all $X \in \mathfrak{g}$
- (2) $\omega(R_{a*}Y) = \text{Ad}(a^{-1})\omega(Y)$ for all $a \in G$, and all tangent vectors Y on E .

If ω is an Ehresmann connection, then the map $\omega(u): P_u \rightarrow \mathfrak{g}$ is onto for every $u \in P$, by (1), so its kernel $H_u = \ker \omega(u)$ is a subspace of P_u having the same dimension as M . This subspace is called the **horizontal subspace** at u (determined by the connection), and tangent vectors in H_u are called **horizontal**. Thus every Ehresmann connection ω on P gives rise to a certain distribution H on P .

4. PROPOSITION. If H is the distribution on P determined by an Ehresmann connection ω , then

- (1) $P_u = V_u \oplus H_u$
- (2) $H_{u \cdot a} = (R_a)_* H_u$
- (3) H is a C^∞ distribution.

Conversely, if H is a distribution on P satisfying (1)–(3), then H is the distribution determined by a unique Ehresmann connection ω .

PROOF. Condition (1) is obvious from the definition of H_u as $\ker \omega(u)$ (and the fact that $\omega(u)$ is onto \mathfrak{g}).

If $Y \in H_u$, then

$$\begin{aligned} \omega(u \cdot a)(R_{a*}Y) &= \omega(R_{a*}Y) \\ &= \text{Ad}(a^{-1})\omega(Y) && \text{by condition (2) in the} \\ &= 0, && \text{definition of a connection} \end{aligned}$$

so $R_{a*}Y \in H_{u \cdot a}$. Since R_{a*} is one-one, and $\dim H_u = \dim H_{u \cdot a}$, it follows that $H_{u \cdot a} = (R_a)_* H_u$.

To prove that H is a C^∞ distribution, choose vector fields $Y_1, \dots, Y_n, \dots, Y_{n+k}$ which span P_v for all v in a neighborhood of u . Let X_1, \dots, X_k be a basis for \mathfrak{g} , so that we can write $\omega = \sum_j \omega^j \cdot X_j$ for ordinary C^∞ 1-forms ω^j on P . Let \bar{Y}_i be the vector field

$$\bar{Y}_i = Y_i - \sum_j \omega^j(Y_i) \sigma(X_j).$$

The C^∞ vector fields \bar{Y}_i are clearly horizontal and span the distribution H in a neighborhood of u .

Conversely, given H , we (must) define ω by $\omega(Y) = 0$ for Y horizontal and $\omega(\sigma(X)) = X$ for $X \in \mathfrak{g}$. Then ω is C^∞ , since $\omega(Y)$ is C^∞ when Y is a horizontal vector field, or when $Y = \sigma(X)$, and these vector fields span the set of all vector fields, over the C^∞ functions. Condition (1) for a connection holds by definition of ω . To prove condition (2), we need only prove it for horizontal Y and vertical Y . If Y is horizontal, then $(R_a)_*Y$ is also, by condition (2) on H , so we have

$$\omega((R_a)_*Y) = 0 = \text{Ad}(a^{-1})\omega(Y).$$

When Y is vertical, we may assume that $Y = \sigma(X)$ for $X \in \mathfrak{g}$. Then $(R_a)_*Y = \sigma(\text{Ad}(a^{-1})X)$ by Proposition 2. So

$$\omega((R_a)_*Y) = \text{Ad}(a^{-1})X = \text{Ad}(a^{-1})\omega(Y),$$

as desired. ♦

Often, a connection is *defined* to be a distribution H satisfying (1)–(3) of Proposition 4, and ω is defined as in the second part of the proof—it is then called the “connection form” for the connection H . Using the decomposition given by (1) of Proposition 4, we can write, for any tangent vector $Y \in P_u$, a unique expression

$$Y = v(Y) + h(Y)$$

where $v(Y)$, the **vertical component** of Y , is vertical and $h(Y)$, the **horizontal component** of Y , is horizontal. As we noted in the proof of Proposition 4,

$$h(Y) = Y - \sum_j \omega^j(Y) \sigma(X_j).$$

From this formula it is clear that $h(Y)$, and hence $v(Y)$, is C^∞ if Y is C^∞ .

From the decomposition $P_u = V_u \oplus H_u$ and the fact that V_u is the kernel of $\pi_*: P_u \rightarrow M_{\pi(u)}$, it is also clear that $\pi_*: H_u \rightarrow M_{\pi(u)}$ is an isomorphism for each $u \in P$. Consequently, for every vector field X on M there is a unique vector field X^* on P such that X^* is everywhere horizontal and $\pi_*(X^*_u) = X_{\pi(u)}$ for all $u \in P$; this vector field X^* is called the **lift** of X . There are two simple propositions about lifts.

5. PROPOSITION. If X is a C^∞ vector field on M , then X^* is a C^∞ vector field on P , and for all $a \in G$ we have $R_{a*}(X^*) = X^*$. Conversely, if Y is a horizontal vector field on P such that $R_{a*}(Y) = Y$ for all $a \in G$, then $Y = X^*$ for a unique vector field X on M .

PROOF. Using local triviality, we can choose a C^∞ vector field X' on some $\pi^{-1}(U) \subset P$ such that $\pi_*(X'_u) = X_{\pi(u)}$ for all $u \in \pi^{-1}(U)$. Then $X^* = h(X')$ is also C^∞ . The other parts are left to the reader. ♦

6. PROPOSITION. If X^* and Y^* are the lifts of vector fields X and Y on M , then

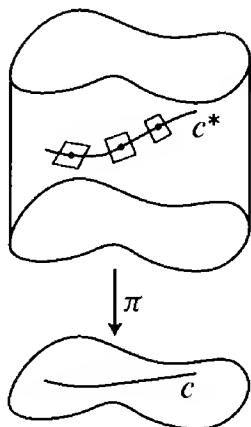
- (1) $X^* + Y^*$ is the lift of $X + Y$
- (2) for every $f: M \rightarrow \mathbb{R}$ we have $(fX)^* = (f \circ \pi) \cdot X^*$
- (3) $h([X^*, Y^*]) = [X, Y]^*$.

PROOF. The first two parts are trivial. For the third, we note that X^* and Y^* are π -related, so by Proposition I.6-3,

$$\pi_*(h[X^*, Y^*]_u) = \pi_*([X^*, Y^*]_u) = [X, Y]_{\pi(u)}. \quad \blacklozenge$$

Our aim now is to develop, and generalize, the material of the last three chapters, from the point of view of Ehresmann connections. The topics to be discussed include parallel translation, covariant derivatives, the curvature and torsion tensors, the structural equations, and the Bianchi identities. Presumably, somewhere along the way an elegant proof of the Test Case will also turn up. Be prepared for some unappetizing complexity—this is the price of invariance! One problem is that all the interesting information about a connection resides in the horizontal subspaces, but formulas about forms on a principal bundle also have to reckon with the vertical vectors [which is what clause (1) in the definition of an Ehresmann connection is designed to do].

We begin with a discussion of parallel translation. A piecewise C^1 curve $\gamma: [0, 1] \rightarrow P$ is called **horizontal** if all tangent vectors $\gamma'(t^+)$ and $\gamma'(t^-)$ are horizontal vectors. Now if $c: [0, 1] \rightarrow M$ is piecewise C^1 , we define a **lift** of c to be a horizontal curve $c^*: [0, 1] \rightarrow P$ such that c^* *covers* c , i.e., such that $\pi \circ c^* = c$. Notice that if c^* is a lift of c , then so is $R_a \circ c^*$, by Proposition 4(2).



7. PROPOSITION. Let $c: [0, 1] \rightarrow M$ be a piecewise C^1 curve, and choose $u_0 \in P$ with $\pi(u_0) = c(0)$. Then there is a unique lift c^* of c with $c^*(0) = u_0$.

PROOF. Using local triviality of the principal bundle, it is easy to show that there is a curve $\gamma: [0, 1] \rightarrow P$ with $\gamma(0) = u_0$ and $\pi \circ \gamma = c$. A lift c^* of c must be of the form $c^*(t) = \gamma(t) \cdot a(t)$ for some $a: [0, 1] \rightarrow G$ with $a(0) = e$. Using the method of proof for Proposition 3 (see also the Remark following it), we find that

$$c^{*'}(t) = R_{a(t)*}(\gamma'(t)) + \sigma(L_{a(t)^{-1}*}a'(t))(c^*(t)).$$

Consequently,

$$\omega(c^{*'}(t)) = \text{Ad}(a(t)^{-1})\omega(\gamma'(t)) + L_{a(t)^{-1}*}a'(t).$$

Now c^* is horizontal if and only if $\omega(c^{*\prime}(t)) = 0$; using the above equation, and remembering the definition of Ad , this means that

$$\begin{aligned} L_{a(t)^{-1}*}a'(t) &= -L_{a(t)^{-1}*}R_{a(t)*}\omega(\gamma'(t)), \\ a'(t) &= -R_{a(t)*}\omega(\gamma'(t)). \end{aligned}$$

In this equation, $\omega(\gamma'(t))$ is a given curve in \mathfrak{g} . If we introduce a coordinate system on G , this equation becomes a differential equation for a , and we know that a unique solution exists locally. So we can always find a lift c^* defined in a neighborhood of any $t \in [0, 1]$, with $c^*(t) = u$ for any given $u \in \pi^{-1}(c(t))$; moreover, this lift is unique.

We now have to show that the lift can be defined on all of $[0, 1]$; clearly we just have to show that a lift on $[0, t_0)$ can be extended past t_0 . To do this, pick a lift \bar{c}^* defined in a neighborhood of t_0 (with any old initial condition $\bar{c}^*(t_0)$). Choose $t_1 < t_0$ so that \bar{c}^* is defined at t_1 , and then choose $a \in G$ with $c^*(t_1) = \bar{c}^*(t_1) \cdot a$. Clearly we can extend c^* past t_0 by letting it be $R_a \circ \bar{c}^*$. ♦

Notice that the last part of this argument is just that used in the proof of Proposition I.5-17. The first part of the argument is much simpler when we are looking for a lift in a neighborhood of a point t with $c'(t) \neq 0$. For then (a portion of) c is the integral curve of a vector field X on M , and c^* is just an integral curve of the lift X^* .

Using Proposition 7, we can now define **parallel translation** of fibres of the principal bundle $\pi: P \rightarrow M$ along any curve $c: [0, 1] \rightarrow M$. For any $u \in \pi^{-1}(c(0))$ we let $\tau_t(u) \in \pi^{-1}(c(t))$ be $c^*(t)$, where c^* is the lift of c with $c^*(0) = u$. In this way we obtain a map

$$\tau_t: \pi^{-1}(c(0)) \rightarrow \pi^{-1}(c(t)).$$

It is clear that $\tau_t \circ R_a = R_a \circ \tau_t$, since $R_a \circ c^*$ is again a lift. The map τ_t is a diffeomorphism whose inverse is just the parallel translation along the reversed portion of c from t to 0.

Consider, in particular, the principal bundle of frames $\pi: F(M) \rightarrow M$. Every frame $u \in F(M)$ determines an isomorphism from \mathbb{R}^n to $M_{\pi(u)}$; namely, we send $e_i \in \mathbb{R}^n$ to $u_i \in M_{\pi(u)}$. This isomorphism will be denoted by the same letter, $u: \mathbb{R}^n \rightarrow M_{\pi(u)}$. It is easy to check that for $\xi \in \mathbb{R}^n$ we have

$$(u \cdot a)(\xi) = u(a \cdot \xi),$$

where the product $a \cdot \xi$ of an $n \times n$ matrix a and a vector $\xi \in \mathbb{R}^n$ is defined in the footnote on page 262. Now suppose $c: [0, 1] \rightarrow M$ is a piecewise C^1 curve,

$X_p \in M_p$ is a tangent vector at $p = c(0)$, and $c^*: [0, 1] \rightarrow F(M)$ is a lift of c with $c^*(0) = u$. There is a unique $\xi \in \mathbb{R}^n$ with $c^*(0)(\xi) = u(\xi) = X_p$; we let

$$\tau_t(X_p) = c^*(t)(\xi),$$

thus defining parallel translation of vectors. To check that this parallel translation is well-defined, we consider any other lift, which must be of the form $\bar{c}^* = R_a \circ c^*$. Then

$$\begin{aligned} X_p &= c^*(0)(\xi) = c^*(0) \cdot a(a^{-1} \cdot \xi) \\ &= \bar{c}^*(0)(a^{-1} \cdot \xi), \end{aligned}$$

so for the parallel translation $\bar{\tau}_t$ defined with respect to \bar{c}^* we have

$$\begin{aligned} \bar{\tau}_t(X_p) &= \bar{c}^*(t)(a^{-1} \cdot \xi) = c^*(t) \cdot a(a^{-1} \cdot \xi) = c^*(t)(\xi) \\ &= \tau_t(X_p). \end{aligned}$$

If we choose $c^*(0) = u$ so that $u((1, 0, \dots, 0)) = X_p$, we see that we can parallel translate X_p by making it the first vector in a basis, parallel translating the basis, and then taking the first vector in the translated basis. It is also easy to see that τ_t is a vector space isomorphism.

Having defined parallel translation of vectors, we can now define covariant differentiation of vector fields by the formula on page 234,

$$(*) \quad \nabla_{X_p} Y = \lim_{h \rightarrow 0} \frac{1}{h} (\tau_h^{-1} Y_{c(h)} - Y_p),$$

where $c(0) = p$ and $c'(0) = X_p$. It is not yet clear that this covariant differentiation is the same as the one we obtained on page 285 from the corresponding Cartan connection $\{s^*\omega\}$; to prove this we will need a Lemma that is also used frequently later on. Given the vector field Y , we consider the function $f_Y: F(M) \rightarrow \mathbb{R}^n$ whose value at a frame v is just the set of components of $Y_{\pi(v)}$ with respect to v —in symbols,

$$f_Y(v) = v^{-1}(Y_{\pi(v)}).$$

8. LEMMA. Let Y be a vector field on M , and let $X_p \in M_p$. Then for any $u \in F(M)$ with $p = \pi(u)$ we have

$$\nabla_{X_p} Y = u(X^*_u(f_Y)),$$

where $X^*_u \in P_u$ is the unique horizontal vector with $\pi_*(X^*_u) = X_p$ [notice that since $f: P \rightarrow \mathbb{R}^n$, the value $X^*_u(f_Y)$ of the vector X^*_u on f_Y is an element of \mathbb{R}^n , so $u(X^*_u(f_Y)) \in M_p$ makes sense].

PROOF. Let c be a curve with $c(0) = p$ and $c'(0) = X_p$, and let c^* be the lift of c with $c^*(0) = u$, so that $c^{*'}(0) = X_u^*$. Recall the definition of the parallel translation $\tau_h^{-1}(Y_{c(h)})$ of $Y_{c(h)}$ along the reversed part of c from 0 to h : we choose $\xi \in \mathbb{R}^n$ with

$$c^*(h)(\xi) = Y_{c(h)} \quad \text{or} \quad c^*(h)^{-1}(Y_{c(h)}) = \xi$$

and then define

$$\tau_h^{-1}(Y_{c(h)}) = c^*(0)(\xi) = u(\xi).$$

Consequently,

$$u \circ c^*(h)^{-1}(Y_{c(h)}) = \tau_h^{-1}(Y_{c(h)}).$$

So we have

$$\begin{aligned} \nabla_{X_p} Y &= \lim_{h \rightarrow 0} \frac{1}{h} [\tau_h^{-1}(Y_{c(h)}) - Y_p] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [u \circ c^*(h)^{-1}(Y_{c(h)}) - u \circ u^{-1}(Y_p)] \\ &= u \left(\lim_{h \rightarrow 0} \frac{1}{h} [c^*(h)^{-1}(Y_{c(h)}) - u^{-1}(Y_p)] \right) \\ &= u \left(\lim_{h \rightarrow 0} \frac{1}{h} [f_Y(c^*(h)) - f_Y(u)] \right) \\ &= u(X_u^*(f_Y)). \quad \blacklozenge \end{aligned}$$

9. PROPOSITION. The ∇ defined by (*) is a Koszul connection; that is,

- (1) $\nabla_{X_p + X'_p} Y = \nabla_{X_p} Y + \nabla_{X'_p} Y$
- (2) $\nabla_{X_p}(Y_1 + Y_2) = \nabla_{X_p} Y_1 + \nabla_{X_p} Y_2$
- (3) $\nabla_{aX_p} Y = a \nabla_{X_p} Y$ for all $a \in \mathbb{R}$
- (4) $\nabla_{X_p}(fY) = f(p) \cdot \nabla_{X_p} Y + X_p(f) \cdot Y_p$
- (5) if X and Y are C^∞ vector fields, then so is $p \mapsto \nabla_{X_p} Y$.

PROOF. Equations (2), (3), and (4) are easy to prove from the definition. Condition (5) follows from Lemma 8, since we can choose a C^∞ section $p \mapsto u(p)$ in a neighborhood of p . Equation (1) also follows from the Lemma, since we have $(X_p + X'_p)^* = X_u^* + X'^*_u$. \blacklozenge

Now let us compare this Koszul connection ∇ with the one obtained from the Cartan connection corresponding to an Ehresmann connection ω on $F(M)$. We can write $\omega = (\omega_j^i)$, where each ω_j^i in the matrix is an ordinary 1-form on $F(M)$. Equivalently, we can write

$$\omega = \sum_{i,j} \omega_j^i \cdot E_i^j,$$

where E_i^j is the matrix with zeros everywhere except for a 1 in the i^{th} row and the j^{th} column, so that

$$(E_i^j)_{\beta}^{\alpha} = \delta_i^{\alpha} \delta_{\beta}^j.$$

Let (x, U) be a coordinate system, and let $s: U \rightarrow F(M)$ be the “natural section”

$$s(q) = \left(\frac{\partial}{\partial x^1} \Big|_q, \dots, \frac{\partial}{\partial x^n} \Big|_q \right).$$

Then the Cartan connection corresponding to ω assigns $(s^* \omega_j^i)$ to this moving frame. So the operation $\bar{\nabla}$ defined on page 285 is determined on U by

$$\bar{\nabla}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \sum_k s^* \omega_j^k \left(\frac{\partial}{\partial x^i} \right) \frac{\partial}{\partial x^k} = \sum_k \omega_j^k \left(s_* \frac{\partial}{\partial x^i} \right) \frac{\partial}{\partial x^k}.$$

10. PROPOSITION. These two connections are the same, $\bar{\nabla} = \nabla$.

PROOF. It obviously suffices to show that $\bar{\nabla}_{\partial/\partial x^i} \partial/\partial x^j = \nabla_{\partial/\partial x^i} \partial/\partial x^j$, since $\bar{\nabla}$ and ∇ are both Koszul connections. Let $f = f_{\partial/\partial x^i}$. By Proposition 8, it suffices to show that

$$\begin{aligned} \sum_k \omega_j^k \left(s_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) \right) \frac{\partial}{\partial x^k} \Big|_p &= s(p) \left(\left(\frac{\partial}{\partial x^i} \right)^*_{s(p)} (f) \right) \\ &= \sum_k \left[\left(\frac{\partial}{\partial x^i} \right)^*_{s(p)} (f^k) \right] \cdot \frac{\partial}{\partial x^k} \Big|_p. \end{aligned}$$

So we need to show that

$$\omega_j^k \left(s_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) \right) = \left(\frac{\partial}{\partial x^i} \right)^*_{s(p)} (f^k).$$

Now, by definition of f we have

$$(*) \quad \sum_k f^k(u) \cdot u_k = \frac{\partial}{\partial x^j} \Big|_{\pi(u)} \quad \text{for all } u = (u_1, \dots, u_n) \in F(M).$$

In particular,

$$f^k(s(q)) = \delta_j^k.$$

Writing equation (*) for $u = s(p) \cdot \exp tN$, and differentiating with respect to t , we obtain

$$\sum_k \frac{d}{dt} \Big|_{t=0} f^k(s(p) \cdot \exp tN) \cdot \frac{\partial}{\partial x^k} \Big|_p + \sum_k f^k(s(p)) \cdot \frac{d}{dt} \Big|_{t=0} (s(p) \cdot \exp tN) = 0,$$

and hence

$$\sum_k \sigma(N)_{s(p)}(f^k) \cdot \frac{\partial}{\partial x^k} \Big|_p = - \sum_k \delta_j^k(s(p) \cdot N)_k = -(s(p) \cdot N)_j.$$

In particular,

$$\begin{aligned} \sum_k \sigma(E_\mu^v)_{s(p)}(f^k) \cdot \frac{\partial}{\partial x^k} \Big|_p &= -(s(p) \cdot E_\mu^v)_j = -(E_\mu^v)_j^\beta \sum_\beta \frac{\partial}{\partial x^\beta} \Big|_p \\ &= -\delta_\mu^\beta \delta_j^v \sum_\beta \frac{\partial}{\partial x^\beta} \Big|_p = -\delta_j^v \frac{\partial}{\partial x^\mu} \Big|_p. \end{aligned}$$

So

$$\sigma(E_\mu^v)_{s(p)}(f^k) = \begin{cases} -\delta_\mu^v & k = \mu \\ 0 & k \neq \mu. \end{cases}$$

Now the lift $(\partial/\partial x^i)^*$ of $\partial/\partial x^i$ at points $s(q)$ is given by

$$\frac{\partial}{\partial x^i}^* = h \left(s_* \left(\frac{\partial}{\partial x^i} \right) \right) = s_* \left(\frac{\partial}{\partial x^i} \right) - \sum_{\mu, v} \omega_v^\mu \left(s_* \left(\frac{\partial}{\partial x^i} \right) \right) \sigma(E_\mu^v).$$

So

$$\begin{aligned} \left(\frac{\partial}{\partial x^i}^* \right)_{s(p)} (f^k) &= \frac{\partial(f^k \circ s)}{\partial x^i}(p) - \sum_{\mu, v} \omega_v^\mu \left(s_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) \right) \sigma(E_\mu^v)_{s(p)}(f^k) \\ &= \frac{\partial(\delta_j^k)}{\partial x^i} - \sum_v \omega_v^k \left(s_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) \right) (-\delta_j^v) \\ &= \omega_j^k \left(s_* \left(\frac{\partial}{\partial x^i} \Big|_p \right) \right) \cdot \diamond \end{aligned}$$

There is no need to repeat here the definition of $\nabla_X A$ for arbitrary tensor fields A ; this can be defined in either of the two ways used in Chapter 6, and we will use any results from that chapter which we require.

We return for a moment to a connection ω on a general principal bundle $\pi: P \rightarrow M$ with group G . Consider a k -form α on P , with values in a vector space V . We define a V -valued $(k+1)$ -form $D\alpha$, the **covariant differential** of α , by

$$D\alpha(Y_1, \dots, Y_{k+1}) = (d\alpha)(hY_1, \dots, hY_{k+1}),$$

where d is the ordinary differential and hY is the horizontal component of Y . In particular, we define the **curvature form** Ω of ω by $\Omega = D\omega$. Thus Ω is a \mathfrak{g} -valued 2-form on P .

11. PROPOSITION. For all $a \in G$, we have $R_a^* \Omega = \text{Ad}(a^{-1})\Omega$. In other words, for all tangent vectors $Y_1, Y_2 \in P_u$ we have

$$R_a^* \Omega(Y_1, Y_2) = \text{Ad}(a^{-1})\Omega(Y_1, Y_2)$$

[this makes sense since $\Omega(Y_1, Y_2) \in \mathfrak{g}$].

PROOF. We have

$$\begin{aligned} R_a^* \Omega(Y_1, Y_2) &= R_a^*(d\omega)(hY_1, hY_2) \\ &= d(R_a^* \omega)(hY_1, hY_2) \\ &= d(\text{Ad}(a^{-1})\omega)(hY_1, hY_2) \\ &= \text{Ad}(a^{-1})[d\omega(hY_1, hY_2)] && \text{since } \text{Ad}(a^{-1}): \mathfrak{g} \rightarrow \mathfrak{g} \\ &&& \text{is linear} \\ &= \text{Ad}(a^{-1})\Omega(Y_1, Y_2). \quad \spadesuit \end{aligned}$$

We will eventually see that this Ω does indeed correspond to the Ω in Chapter 7, but we first introduce the analogue of Θ . This analogue cannot be defined for connections in all principal bundles, but only for connections ω in the bundle of frames $\pi: F(M) \rightarrow M$. On this bundle we have a certain \mathbb{R}^n -valued 1-form θ , defined by

$$\theta_u(Y_u) = u^{-1}(\pi_* Y_u).$$

We will call this 1-form the **canonical form** or the **dual form** of the principal bundle $F(M)$. To see the appropriateness of the latter term, consider a section $s: U \rightarrow F(M)$ given by $s = (X_1, \dots, X_n)$. For any tangent vector $Y_p \in M_p$ we have

$$s^* \theta(Y_p) = \theta_{s(p)}(s_* Y_p) = s(p)^{-1}(Y_p),$$

so for the i^{th} component θ^i of θ we have

$$\begin{aligned} s^*\theta^i(Y_p) &= i^{\text{th}} \text{ component of } s(p)^{-1}(Y_p) \\ &= i^{\text{th}} \text{ component of } Y_p \text{ with respect to the basis } X_1(p), \dots, X_n(p). \end{aligned}$$

Thus, the $s^*\theta^i$ are just the dual forms for the moving frame (X_1, \dots, X_n) .

We now define the **torsion form** Θ of a connection ω on $F(M)$ by $\Theta = D\theta$ (this depends on ω , since h , and hence D , depends on ω).

12. PROPOSITION. For all $A \in \text{GL}(n, \mathbb{R})$ we have

$$R_A^*\theta = A^{-1} \cdot \theta, \quad R_A^*\Theta = A^{-1} \cdot \Theta.$$

In other words, for all tangent vectors $Y_1, Y_2 \in P_u$ we have

$$\begin{aligned} R_A^*\theta(Y_1) &= A^{-1} \cdot \theta(Y_1) \\ R_A^*\Theta(Y_1, Y_2) &= A^{-1} \cdot \Theta(Y_1, Y_2) \end{aligned}$$

[these equations make sense since $\theta(Y_1), \Theta(Y_1, Y_2) \in \mathbb{R}^n$].

PROOF. For θ we have

$$\begin{aligned} R_A^*\theta(Y_1) &= \theta_{u \cdot a}(R_{A*}Y_1) = (u \cdot A)^{-1}(\pi_*Y_1) \\ &= A^{-1} \cdot u^{-1}(\pi_*Y_1) \\ &= A^{-1} \cdot \theta(Y_1). \end{aligned}$$

Then for Θ we have

$$\begin{aligned} R_A^*\Theta(Y_1, Y_2) &= R_A^*(d\theta)(hY_1, hY_2) \\ &= d(R_A^*\theta)(hY_1, hY_2) \\ &= d(A^{-1} \cdot \theta)(hY_1, hY_2) \\ &= A^{-1} \cdot d\theta(hY_1, hY_2) \\ &= A^{-1} \cdot \Theta(Y_1, Y_2). \quad \spadesuit \end{aligned}$$

A connection ω in the bundle of frames $F(M)$ also allows us to define certain special vector fields in $F(M)$. For $\xi \in \mathbb{R}^n$ we define the **basic vector field** $B(\xi)$ **corresponding to** ξ by letting $B(\xi)_u$ be the unique horizontal vector at u such that $\pi_*(B(\xi)_u) = u(\xi)$. In particular, $B(e_i)_u$ is the unique horizontal vector at u which covers u_i .

13. PROPOSITION. For all $\xi \in \mathbb{R}^n$ we have

- (1) $\theta(B(\xi)) = \xi$
- (2) $R_{A*}B(\xi) = B(A^{-1} \cdot \xi)$ for all $A \in \text{GL}(n, \mathbb{R})$.

Moreover, if $\xi \neq 0$, then $B(\xi)$ is nowhere 0. Consequently, if ξ_1, \dots, ξ_n is a basis for \mathbb{R}^n , then $B(\xi_1)_u, \dots, B(\xi_n)_u$ is a basis for H_u .

PROOF. (1) and (2) are left to the reader. For the third assertion, note that if $B(\xi)_u = 0$, then

$$0 = \pi_*(B(\xi)_u) = u(\xi),$$

so $\xi = 0$. The final assertion follows immediately. ♦

To prove the structural equations in our new setup, we need two lemmas; the first involves fundamental vector fields and basic vector fields, but the second, which holds for connections in any bundle, involves fundamental vector fields and arbitrary horizontal vector fields.

14. LEMMA. Consider the basic vector fields determined by a connection on the bundle of frames $F(M)$. For every $N \in \mathfrak{gl}(n, \mathbb{R})$ and $\xi \in \mathbb{R}^n$ we have

$$[\sigma(N), B(\xi)] = B(N \cdot \xi).$$

PROOF. Since $\phi_t(u) = u \cdot \exp tN = R_{\exp tN}(u)$ is the 1-parameter group of diffeomorphisms generated by $\sigma(N)$, we have

$$\begin{aligned} [\sigma(N), B(\xi)] &= \lim_{t \rightarrow 0} \frac{1}{t} [B(\xi) - R_{\exp tN*}B(\xi)] \\ &= \lim_{t \rightarrow 0} \frac{1}{t} [B(\xi) - B([\exp -tN] \cdot \xi)] && \text{by Proposition 13} \\ &= B\left(\lim_{t \rightarrow 0} \frac{1}{t} (\xi - (\exp -tN) \cdot \xi)\right) && \text{since } \xi \mapsto B(\xi) \text{ is linear onto } H_u \\ &= B(N \cdot \xi). \quad \diamond \end{aligned}$$

15. LEMMA. Consider a connection on any principal bundle P over M with group G . For any $X \in \mathfrak{g}$ and any horizontal vector field Y on P , the vector field $[\sigma(X), Y]$ is also horizontal.

PROOF. We have

$$[\sigma(X), Y] = \lim_{t \rightarrow 0} \frac{1}{t} [Y - R_{\exp tX*}(Y)].$$

Since Y is horizontal, so is each $R_{\exp tX*}(Y)$. ♦

16. THEOREM. Let ω be a connection on a principal bundle P over M . If P is the bundle of frames, with the dual form θ , and the torsion form Θ determined by ω , then we have the **first structural equation**:

$$d\theta(Y_1, Y_2) = -\{\omega(Y_1) \cdot \theta(Y_2) - \omega(Y_2) \cdot \theta(Y_1)\} + \Theta(Y_1, Y_2) \quad \text{for all } Y_1, Y_2 \in P_u$$

[where $\omega(Y_1) \cdot \theta(Y_2)$ is the action of the matrix $\omega(Y_1)$ on $\theta(Y_2) \in \mathbb{R}^n$].

If P is any principal bundle, and Ω is the curvature form of ω , then we have the **second structural equation**:

$$d\omega(Y_1, Y_2) = -[\omega(Y_1), \omega(Y_2)] + \Omega(Y_1, Y_2) \quad \text{for all } Y_1, Y_2 \in P_u.$$

PROOF. Since each Y_i is the sum of a vertical and a horizontal vector, and since both sides of the first structural equation are skew-symmetric and bilinear, we can prove this equation by considering 3 cases.

Case 1. Y_1 and Y_2 are horizontal. Then $\omega(Y_i) = 0$, so the equation reduces to the definition of Θ as $D\theta$.

Case 2. Y_1 and Y_2 are vertical. Then the right side is 0. If we extend Y_1 and Y_2 to vertical vector fields \tilde{Y}_1 and \tilde{Y}_2 , then the left side is the value at u of

$$Y_1(\theta(Y_2)) - Y_2(\theta(Y_1)) - \theta([Y_1, Y_2]),$$

which is 0, since $[Y_1, Y_2]$ is also vertical.

Case 3. Y_1 is vertical and Y_2 is horizontal. Let $Y_1 = \sigma(N)_u$ for $N \in \mathfrak{gl}(n, \mathbb{R})$ and let $Y_2 = B(\xi)_u$ for $\xi \in \mathbb{R}^n$. Then

$$-\{\omega(Y_1) \cdot \theta(Y_2) - \omega(Y_2) \cdot \theta(Y_1)\} + \Theta(Y_1, Y_2) = -N \cdot \xi + 0 + 0,$$

while we have

$$\begin{aligned} d\theta(Y_1, Y_2) &= \sigma(N)(\theta(B(\xi)))(u) - B(\xi)(\theta(\sigma(N)))(u) - \theta([\sigma(N), B(\xi)])(u) \\ &= 0 - 0 - \theta(B(N \cdot \xi))(u) \quad \text{by Lemma 14} \\ &= -N \cdot \xi. \end{aligned}$$

This proves the first structural equation.

The second structural equation will be proved similarly.

Case 1. Y_1 and Y_2 are horizontal. The proof is as before.

Case 2. Y_1 and Y_2 are vertical. Let $Y_i = \sigma(X_i)_u$ for $X_i \in \mathfrak{g}$. Then $\Omega(Y_1, Y_2) = 0$, while

$$\begin{aligned} d\omega(Y_1, Y_2) &= \sigma(X_1)(\omega(\sigma(X_2)))(u) \\ &\quad - \sigma(X_2)(\omega(\sigma(X_1)))(u) - \omega([\sigma(X_1), \sigma(X_2)])(u) \\ &= 0 - 0 - \omega(\sigma([X_1, X_2]))(u) \quad \text{by Proposition 2} \\ &= -[X_1, X_2] = -[\omega(Y_1), \omega(Y_2)]. \end{aligned}$$

Case 3. Y_1 is vertical and Y_2 is horizontal. Then the right side is 0. If we extend Y_2 to a horizontal vector field \tilde{Y}_2 and let $Y_1 = \sigma(X)_u$ for $X \in \mathfrak{g}$, then the left side is the value at u of

$$\sigma(X)(\omega(\tilde{Y}_2)) - \tilde{Y}_2(\omega(\sigma(X))) - \omega([\sigma(X), \tilde{Y}_2]) = 0, \quad \text{by Lemma 15. } \spadesuit$$

The structural equations for a connection ω on the bundle of frames $F(M)$ can also be written in terms of ordinary forms. With respect to the standard basis e_1, \dots, e_n of \mathbb{R}^n we can write the \mathbb{R}^n -valued forms θ and Θ as

$$\theta = \sum_i \theta^i \cdot e_i \quad \Theta = \sum_i \Theta^i \cdot e_i,$$

for certain ordinary forms θ^i and Θ^i . Similarly, with respect to the basis E_i^j of $\mathfrak{gl}(n, \mathbb{R})$ introduced previously, we can write

$$\omega = \sum_{i,j} \omega_j^i \cdot E_i^j \quad \Omega = \sum_{i,j} \Omega_j^i \cdot E_i^j.$$

It is easy to see that the structural equations can then be written

$$\begin{aligned} d\theta^i &= -\sum_j \omega_j^i \wedge \theta^j + \Theta^i \\ d\omega_j^i &= -\sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i. \end{aligned}$$

Instead of introducing these forms explicitly, it will be convenient to write the structural equations in the abbreviated form

$$\begin{aligned} d\theta &= -\omega \wedge \theta + \Theta \\ d\omega &= -\omega \wedge \omega + \Omega. \end{aligned}$$

For a section $s: U \rightarrow F(M)$ we obtain

$$\begin{aligned} d(s^*\theta^i) &= -\sum_j s^*\omega_j^i \wedge s^*\theta^j + s^*\Theta^i \\ d(s^*\omega_j^i) &= -\sum_k s^*\omega_k^i \wedge s^*\omega_j^k + s^*\Omega_j^i. \end{aligned}$$

Since $s^*\theta^i$ are just the dual forms to the moving frame s , this shows that $s^*\Theta^i$ and $s^*\Omega_j^i$ are precisely the torsion and curvature forms* for the moving frame s which are determined by the Cartan connection which assigns $s^*\omega$ to s . Propositions 11 and 12 correspond to Proposition 15 of Chapter 7. Recall that the formulas of this proposition allowed us to define the tensors T and R . Essentially equivalent, but much neater, definitions will now be given directly from Propositions 11 and 12.

Let ω be a connection on $F(M)$ with torsion form Θ and curvature form Ω . For $X_1, X_2 \in M_p$, we let

$$T(X_1, X_2) = u(\Theta(\bar{X}_1, \bar{X}_2)) \in M_p,$$

where $\bar{X}_i \in F(M)_u$ are any vectors with $\pi_*(\bar{X}_i) = X_i$. This definition is inde-

*In Ehresmann's original treatment, the torsion and curvature forms Θ and Ω on $F(M)$ were defined by the structural equations. The definition in terms of D was given by Ambrose and Singer, *A Theorem on Holonomy*, Trans. Amer. Math. Soc. **75** (1953), 428–443. This paper also introduced much of the convenient terminology, like “horizontal vectors”.

pendent of the choice of u , and of $\bar{X}_i \in F(M)_u$. To see this, first note that if \bar{X}_1 is replaced by $\bar{\bar{X}}_1$ with $\pi_*(\bar{\bar{X}}_1) = X_1$, then $\bar{X}_1 - \bar{\bar{X}}_1$ is vertical, so

$$\Theta(\bar{X}_1 - \bar{\bar{X}}_1, \bar{X}_2) = D\theta(\bar{X}_1 - \bar{\bar{X}}_1, \bar{X}_2) = 0;$$

similarly, \bar{X}_2 may be replaced by any $\bar{\bar{X}}_2$ with $\pi_*(\bar{\bar{X}}_2) = X_2$, without changing the value of $u(\Theta(\bar{X}_1, \bar{X}_2))$. Then note that if we change u to $u \cdot A$, we can pick $R_{A*}\bar{X}_i$ for the new \bar{X}_i , and we have

$$\begin{aligned} (u \cdot A)(\Theta(R_{A*}\bar{X}_1, R_{A*}\bar{X}_2)) &= (u \cdot A)(A^{-1} \cdot \Theta(\bar{X}_1, \bar{X}_2)) && \text{by Proposition 12} \\ &= u(\Theta(\bar{X}_1, \bar{X}_2)). \end{aligned}$$

Since Θ is a form, it is clear that T is skew-symmetric.

Similarly, for $X_1, X_2, X_3 \in M_p$ we let

$$R(X_1, X_2)X_3 = u(\Omega(\bar{X}_1, \bar{X}_2) \cdot (u^{-1}X_3));$$

here $\Omega(\bar{X}_1, \bar{X}_2) \in \mathfrak{gl}(n, \mathbb{R})$ is an $n \times n$ matrix, so it acts on $u^{-1}X_3 \in \mathbb{R}^n$. Just as before, we see that the definition does not depend on the choice of \bar{X}_1 or \bar{X}_2 . If we change u to $u \cdot A$, then by Proposition 11,

$$\begin{aligned} (u \cdot A)(\Omega(R_{A*}\bar{X}_1, R_{A*}\bar{X}_2) \cdot ([u \cdot A]^{-1}X_3)) \\ &= (u \cdot A)([A^{-1}\Omega(\bar{X}_1, \bar{X}_2)A] \cdot (A^{-1} \cdot u^{-1}(X_3))) \\ &= u(\Omega(\bar{X}_1, \bar{X}_2) \cdot (u^{-1}X_3)). \end{aligned}$$

Since Ω is a 2-form, it is clear that R is skew-symmetric in X_1 and X_2 .

In Chapter 7, we mentioned that the structural equations could be used to prove that the torsion and curvature tensors T and R , defined for a Cartan connection in terms of Θ^i and Ω_j^i , were just those derived from the ∇ which could also be defined for the Cartan connection. Here we will actually carry out this proof for an Ehresmann connection.

17. PROPOSITION. For any vector fields X, Y, Z on M we have

$$\begin{aligned} T(X, Y) &= \nabla_X Y - \nabla_Y X - [X, Y] \\ R(X, Y)Z &= \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X, Y]} Z. \end{aligned}$$

PROOF. Recall that we defined $f_Y: F(M) \rightarrow \mathbb{R}^n$ by $f_Y(u) = u^{-1}(Y_{\pi(u)})$. If Y^* is the lift of Y , we can write $f_Y(u) = \theta(Y^*_u)$. So Lemma 8 gives

$$\nabla_X Y(p) = u(X^*_u(\theta(Y^*))) \quad \text{for } u \in \pi^{-1}(p).$$

Consequently,

$$\begin{aligned} T(X_p, Y_p) &= u(\Theta(X^*_u, Y^*_u)) \\ &= u(D\theta(X^*_u, Y^*_u)) \\ &= u(d\theta(X^*_u, Y^*_u)) \\ &= u(X^*_u(\theta(Y^*)) - Y^*_u(\theta(X^*)) - \theta([X^*, Y^*](u))) \\ &= \nabla_{X_p} Y - \nabla_{Y_p} X - [X, Y]_p, \end{aligned}$$

since $\pi_*([X^*, Y^*]) = [X, Y]$.

To prove the second equality, we note that since X^* and Y^* are horizontal, we have

$$\begin{aligned} \Omega(X^*_u, Y^*_u) &= d\omega(X^*, Y^*)(u) \\ &= X^*(\omega(Y^*))(u) - Y^*(\omega(X^*))(u) - \omega([X^*, Y^*])(u) \\ &= -\omega([X^*, Y^*])(u). \end{aligned}$$

If we set the vertical component of $[X^*, Y^*]$ equal to

$$v[X^*, Y^*](u) = \sigma(N)_u \quad \text{for } N \in \mathfrak{gl}(n, \mathbb{R}),$$

then we obtain

$$\Omega(X^*_u, Y^*_u) = -N,$$

so

$$\begin{aligned} (1) \quad R(X_p, Y_p)Z_p &= u(\Omega(X^*_u, Y^*_u) \cdot (u^{-1}Z_p)) \\ &= u(-N \cdot f_Z(u)). \end{aligned}$$

On the other hand, since we also have $f_Z(u) = \theta(Z^*_u)$, we obtain

$$\begin{aligned}
(2) \quad & \nabla_X \nabla_Y Z(p) - \nabla_Y \nabla_X Z(p) - \nabla_{[X, Y]} Z(p) \\
&= u(X^*_u(Y^* f_Z) - Y^*_u(X^* f_Z) - (h[X^*, Y^*]_u) f_Z) \\
&= u((v[X^*, Y^*]_u) f_Z).
\end{aligned}$$

The expressions in (1) and (2) are equal, since

$$\begin{aligned}
\sigma(N)_u f_Z &= \lim_{t \rightarrow 0} \frac{1}{t} [f_Z(u \cdot \exp tN) - f_Z(u)] \\
&= \lim_{t \rightarrow 0} \frac{1}{t} [(\exp tN)^{-1} \cdot f_Z(u) - f_Z(u)] \\
&= -N \cdot f_Z(u). \quad \blacklozenge
\end{aligned}$$

One of the steps in this proof is of sufficient importance to be stated explicitly, along with a counterpart, in the following corollary of the structural equations.

18. PROPOSITION. If ω is a connection on the bundle of frames $F(M)$, and B_1, B_2 are basic vector fields, then the horizontal component of $[B_1, B_2](u)$ is the value of $B(-\Theta(B_{1u}, B_{2u}))$ at u .

If ω is a connection on any principal bundle P over M , and Y_1, Y_2 are horizontal vector fields, then the vertical component of $[Y_1, Y_2](u)$ is the value of $\sigma(-\Omega(Y_{1u}, Y_{2u}))$ at u .

PROOF. Since B_1 and B_2 are horizontal, the first structural equation gives

$$\begin{aligned}
\Theta(B_{1u}, B_{2u}) &= d\theta(B_1, B_2)(u) \\
&= B_1(\theta(B_2))(u) - B_2(\theta(B_1))(u) - \theta([B_1, B_2])(u) \\
&= -\theta(h[B_1, B_2](u)) \quad \text{since } \theta = 0 \text{ on vertical vectors.}
\end{aligned}$$

Since we can always write $h[B_1, B_2](u)$ as $B(\xi)_u$ for some $\xi \in \mathbb{R}^n$, the first result follows from Proposition 13(1).

Since Y_1 and Y_2 are horizontal, the second structural equation gives

$$\Omega(Y_{1u}, Y_{2u}) = d\omega(Y_1, Y_2)(u) = -\omega([Y_1, Y_2](u)),$$

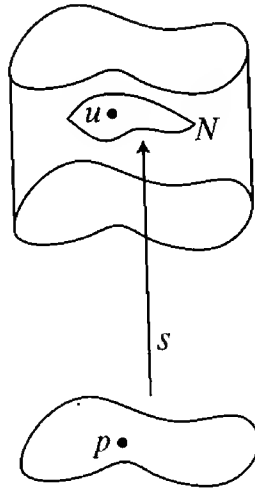
which gives the second result. \blacklozenge

Suddenly, we are ready for

19. **THEOREM (THE TEST CASE; SEVENTH VERSION).** Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold for which the curvature tensor R (for the Levi-Civita connection) is 0. Then M is locally isometric to \mathbb{R}^n with its usual Riemannian metric.

PROOF. On the bundle of frames $F(M)$ we have a connection ω with $\Theta = 0$ and $\Omega = 0$, for which parallel translation preserves the inner product $\langle \cdot, \cdot \rangle$.

Step 1. The second part of Proposition 18 shows that the bracket of two horizontal vector fields is again horizontal. Thus, *the distribution H is integrable*. At a point $p \in M$ choose an *orthonormal* frame $u \in \pi^{-1}(p)$, and let N be the (n -dimensional) integral manifold of H through u . Clearly, N is locally the image of a section $s: U \rightarrow F(M)$ with $s(p) = u$.



Step 2. Consider the basic vector fields $B(e_i)$, for e_1, \dots, e_n the standard basis of \mathbb{R}^n . Since they are horizontal, the bracket $[B(e_i), B(e_j)]$ is horizontal. But the first part of Proposition 18 shows that the horizontal component is 0, so $[B(e_i), B(e_j)] = 0$. If we let $X_i(q) = \pi_*(B(e_i)(s(q)))$, then we also have $[X_i, X_j] = 0$, since $B(e_i)$ and X_i are π -related. But, by definition of $B(\xi)$, we have

$$\begin{aligned} \pi_*(B(e_i)(s(q))) &= s(q)(e_i) \\ &= i^{\text{th}} \text{ vector of the frame } s(q). \end{aligned}$$

Thus, $s = (X_1, \dots, X_n)$ where $[X_i, X_j] = 0$; hence there is a coordinate system x^1, \dots, x^n with $X_i = \partial/\partial x^i$.

Step 3. We claim this is the desired coordinate system. We just have to show that $s(q)$ is always orthonormal, so it suffices to show that $s(q)$ is the parallel translate of $u = s(p)$ along any curve c in U from p to q . This is obvious: to translate u along c , we choose a lift c^* of c with $c^*(0) = u$, and then the translate of u is $c^*(1)$; but clearly $c^* = s \circ c$. ♦

To wrap things up, we present the new version of the Bianchi identities.

20. THEOREM. For a connection ω on the bundle of frames $F(M)$, with dual form θ , torsion form Θ , and curvature form Ω , we have

- (1) (Bianchi's first identity) $D\Theta = \Omega \wedge \theta$. In other words, for $X, Y, Z \in F(M)_u$ we have

$$\begin{aligned} D\Theta(X, Y, Z) &= \frac{(2+1)!}{2!1!} \cdot \frac{1}{3!} \cdot [\Omega(X, Y) \cdot \theta(Z) - \Omega(Y, X) \cdot \theta(Z) \\ &\quad + \Omega(Y, Z) \cdot \theta(X) - \Omega(Z, Y) \cdot \theta(X) \\ &\quad + \Omega(Z, X) \cdot \theta(Y) - \Omega(X, Z) \cdot \theta(Y)] \\ &= \Omega(X, Y) \cdot \theta(Z) + \Omega(Y, Z) \cdot \theta(X) + \Omega(Z, X) \cdot \theta(Y). \end{aligned}$$

For a connection ω on any principal bundle, with curvature form Ω , we have

- (2) (Bianchi's second identity) $D\Omega = 0$.

PROOF. Applying d to the first structural equation, $d\theta = -\omega \wedge \theta + \Theta$, we obtain

$$0 = -(d\omega \wedge \theta) + (\omega \wedge d\theta) + d\Theta.$$

So

$$\begin{aligned} D\Theta(X, Y, Z) &= d\Theta(hX, hY, hZ) \\ &= (d\omega \wedge \theta)(hX, hY, hZ) - 0 \\ &= (\Omega \wedge \theta)(X, Y, Z), \end{aligned}$$

since $d\omega(hA, hB) = \Omega(A, B)$ and $\theta(hA) = \theta(A)$ for all $A, B \in F(M)_u$.

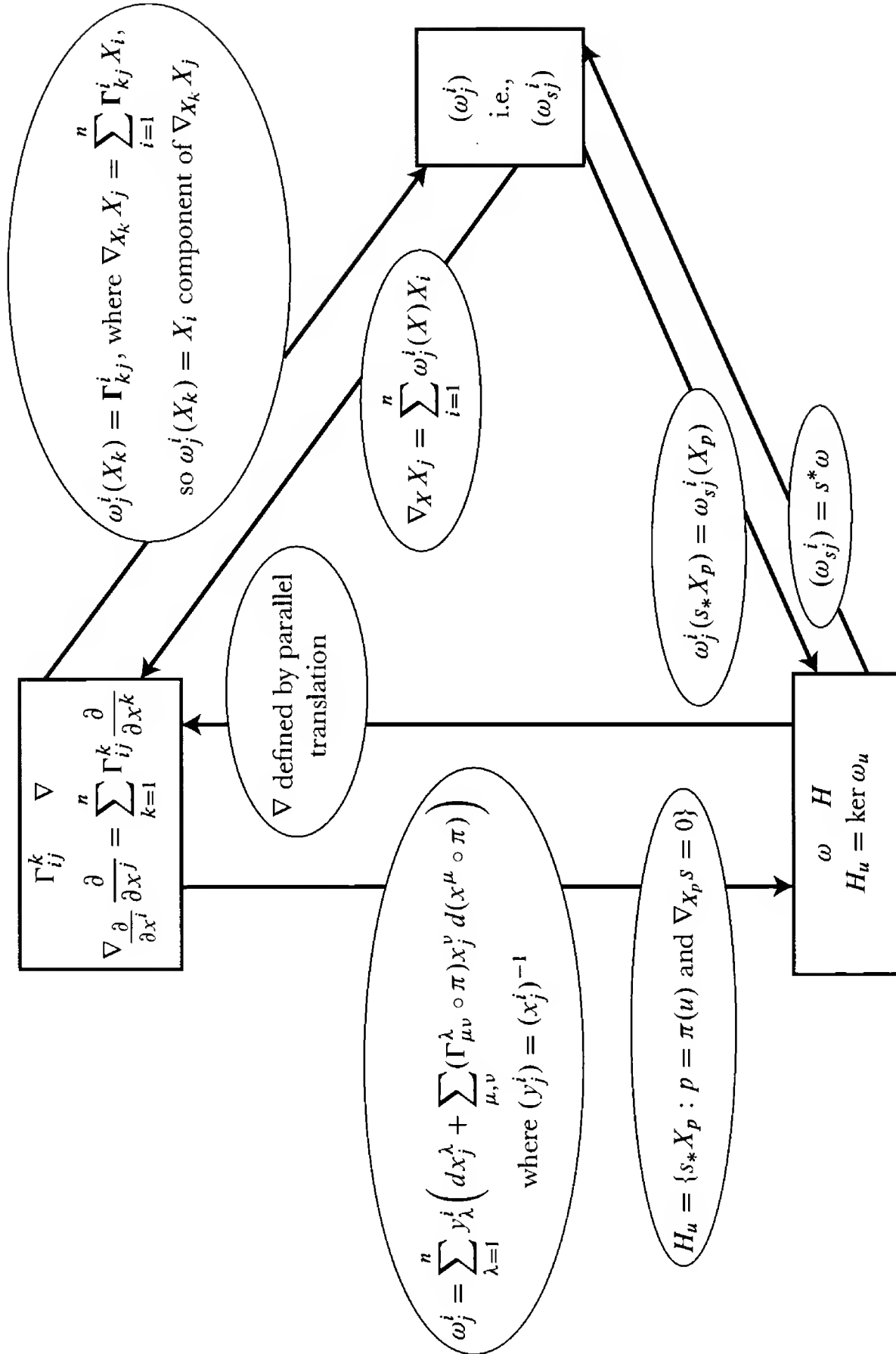
Applying d to the second structural equation, $d\omega = -\omega \wedge \omega + \Omega$, we obtain

$$0 = -(d\omega \wedge \omega) + (\omega \wedge d\omega) + d\Omega.$$

So

$$\begin{aligned}
 D\Omega(X, Y, Z) &= d\Omega(hX, hY, hZ) \\
 &= (d\omega \wedge \omega)(hX, hY, hZ) - (\omega \wedge d\omega)(hX, hY, hZ) \\
 &= 0,
 \end{aligned}$$

since $\omega(hA) = 0$ for all $A \in F(M)_u$ [but recall that $(d\omega \wedge \omega) - (\omega \wedge d\omega)$ is not itself 0, since matrix multiplication is not commutative]. ♦



SUMMARY

The diagram on the opposite page summarizes the relationship between the various definitions of a connection:

- (1) n^3 functions Γ_{ij}^k assigned to each coordinate system (classical),
- (2) ∇ operator on vector fields (Koszul),
- (3) $n \times n$ matrix of 1-forms (ω_j^i) assigned to each moving frame (É. Cartan),
- (4) $n \times n$ matrix-valued 1-form ω on $F(M)$ (Ehresmann),
- (5) distribution H on $F(M)$.

The relationship between (1) and (2) is immediate, as is the relationship between (4) and (5). In Chapter 7 we saw how to pass between (1)-(2) and (3), and in this chapter we saw how to pass between (4)-(5) and (3). We can also go directly from (5) to (1), since ∇ is defined by parallel translation. It remains to indicate how one passes directly from (1)-(2) to (5).

21. PROPOSITION. Let ∇ be a Koszul connection on M . For $u \in F(M)$, let

$$H_u = \{s_* X_p : p = \pi(u) \text{ and } \nabla_{X_p} s = 0\}$$

[where $\nabla_{X_p} s$ denotes $(\nabla_{X_p} X_1, \dots, \nabla_{X_p} X_n)$ if $s = (X_1, \dots, X_n)$]. Then H_u is a subspace of P_u , and the distribution $u \mapsto H_u$ defines a connection on $F(M)$ for which the covariant derivative is the given ∇ .

PROOF. Suppose that $s \cdot a$ is another section with $a(p) = I$ and $\nabla_{X_p}(s \cdot a) = 0$. Then

$$\begin{aligned} (1) \quad 0 &= \nabla_{X_p}(s \cdot a) = s \cdot X_p(a) + (\nabla_{X_p} s) \cdot a \\ &= s \cdot X_p(a), \end{aligned}$$

which implies that $X_p(a)$ is the zero matrix. So by Proposition 3 we have

$$\begin{aligned} (s \cdot a)_*(X_p) &= s_* X_p + \sigma(X_p(a))(s(p)) \\ &= s_* X_p. \end{aligned}$$

Thus H_u contains exactly one vector X_u^* with $\pi_*(X_u^*) = X_p$. Now given vectors $A_u, B_u \in H_u$, let $\pi_* A_u = X_p$ and $\pi_* B_u = Y_p$. We can choose s with $s(p) = u$ and $\nabla_{X_p} s = \nabla_{Y_p} s = 0$; then we also have $\nabla_{X_p + Y_p} s = 0$, and we must have

$$A_u + B_u = s_* X_p + s_* Y_p = s_*(X_p + Y_p) \in H_u.$$

Thus H_u is a subspace.* It has the same dimension as M , and contains no vertical vectors except 0, so we have $F(M)_u = V_u \oplus H_u$. It is clear that $R_{A*}H_u = H_{u \cdot A}$ for all $A \in \text{GL}(n, \mathbb{R})$, since $s \cdot A$ also satisfies $\nabla_{X_p} s \cdot A = 0$.

To prove that H is a C^∞ distribution, we consider a C^∞ vector field X on M , and a C^∞ section s . We claim that there is a function $a: U \rightarrow \text{GL}(n, \mathbb{R})$, defined in some open set U , such that $\nabla_{X_p}(s \cdot a) = 0$ for all $p \in U$. This is because equation (1) shows that this condition is equivalent to

$$0 = s \cdot X_p(a) + (\nabla_{X_p} s) \cdot a,$$

and this is a differential equation for a . For each $A \in \text{GL}(n, \mathbb{R})$ and $p \in U$, let $a_{p,A}$ be the solution with $a_{p,A}(p) = s(p) \cdot A$. For every $u \in F(M)$ we have $u = s(\pi(u)) \cdot A(u)$ for some $A(u) \in \text{GL}(n, \mathbb{R})$ which depends differentiably on u . Then the vector field

$$X^*(u) = [s \cdot a_{\pi(u), A(u)}]_*(X_{\pi(u)}) \in F(M)_u$$

is C^∞ , since $a_{p,A}$ is C^∞ in p and A .

Thus the distribution H is a connection on $F(M)$. For a curve c in M , consider a curve c^* in $F(M)$ with $\pi \circ c^* = c$ and all $c^{*'}(t)$ horizontal. By definition of the connection, this means that for some $X_p \in M_p = M_{c(t)}$ we have

$$c^{*'}(t) = s_*(X_p) \quad \text{where} \quad \nabla_{X_p} s = 0.$$

Of course, X_p must be $\pi_* s_* X_p = \pi_* c^{*'}(t) = c'(t)$. So $\nabla_{c'(t)} s = 0$. From this it is easy to see that the components of c^* are vector fields along c which are parallel along c , with respect to the original ∇ . Hence parallel translation defined with respect to the connection H is the same as parallel translation for ∇ . This implies that the covariant derivatives are also the same. ♦

Proposition 21 gives the most geometric way of going from ∇ to H , and hence to ω . We can also give a computational description of ω , in terms of a coordinate system (x, U) on M . Recall that $x_\# = (x^i \circ \pi, x_j^i)$ is a coordinate system on $\pi^{-1}(U)$, where we write a frame u as

$$u_j = \sum_{i=1}^n x_j^i(u) \cdot \frac{\partial}{\partial x^i} \Big|_{\pi(u)}.$$

At each u , the non-singular matrix $(x_j^i(u))$ has an inverse matrix $(y_j^i(u)) = (x_j^i(u))^{-1}$.

* We can also define H_u simply as $s_*(M_p)$, where s is a section with $s(p) = u$ and $(\nabla s)(p) = 0$.

22. PROPOSITION. Let Γ_{ij}^k be the n^3 functions assigned to the coordinate system (x, U) by a classical connection on M . For the corresponding Ehresmann connection $\omega = (\omega_j^i) = \sum_{i,j} \omega_j^i \cdot E_i^j$ we have

$$\omega_j^i = \sum_{\lambda=1}^n y_\lambda^i \left(dx_j^\lambda + \sum_{\mu, \nu} (\Gamma_{\mu\nu}^\lambda \circ \pi) x_j^\nu d(x^\mu \circ \pi) \right) \quad \text{on } \pi^{-1}(U).$$

PROOF. Let $s = (\partial/\partial x^1, \dots, \partial/\partial x^n)$ be the natural section. We know that the matrix of 1-forms (ω_{sj}^i) for the corresponding Cartan connection is given by

$$\Gamma_{\mu j}^i = (\omega_{sj}^i) \left(\frac{\partial}{\partial x^\mu} \right),$$

and consequently the corresponding Ehresmann connection satisfies

$$(1) \quad \Gamma_{\mu j}^i(p) = \omega_j^i \left(s_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) \right).$$

If we write

$$s_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) = \sum_i a^i \frac{\partial}{\partial (x^i \circ \pi)} \Big|_{s(p)} + \sum_{i,j} b_j^i \frac{\partial}{\partial x_j^i} \Big|_{s(p)},$$

then

$$\begin{aligned} a^i &= s_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) (x^i \circ \pi) = \frac{\partial (x^i \circ \pi \circ s)}{\partial x^\mu} \Big|_p = \frac{\partial x^i}{\partial x^\mu} \Big|_p = \delta_\mu^i \\ b_j^i &= s_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) (x_j^i) = \frac{\partial (x_j^i \circ s)}{\partial x^\mu} \Big|_p = \frac{\partial \delta_j^i}{\partial x^\mu} \Big|_p = 0, \end{aligned}$$

so

$$s_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) = \frac{\partial}{\partial (x^\mu \circ \pi)} \Big|_{s(p)}.$$

A similar calculation shows that for any $A \in \text{GL}(n, \mathbb{R})$ we have

$$(2) \quad \frac{\partial}{\partial (x^\mu \circ \pi)} \Big|_{s(p) \cdot A} = (s \cdot A)_* \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) = R_{A*} \left(\frac{\partial}{\partial x^\mu} \Big|_p \right).$$

Now any $u \in \pi^{-1}(U)$ can be written $u = s(\pi(u)) \cdot (x_j^i(u)) = s(p) \cdot A$, say. So for the matrix (ω_j^i) we have

$$\begin{aligned} (\omega_j^i) \left(\frac{\partial}{\partial(x^\mu \circ \pi)} \Big|_u \right) &= (\omega_j^i) \left(R_{A*} \frac{\partial}{\partial x^\mu} \Big|_p \right) && \text{by (2)} \\ &= A^{-1} (\omega_j^i) \left(\frac{\partial}{\partial x^\mu} \Big|_p \right) A && \text{by definition of an} \\ & && \text{Ehresmann connection} \\ &= A^{-1} (\Gamma_{\mu j}^i(p)) A && \text{by (1),} \end{aligned}$$

so that

$$\begin{aligned} \omega_j^i \left(\frac{\partial}{\partial(x^\mu \circ \pi)} \Big|_u \right) &= \sum_{\lambda, v} (A^{-1})_\lambda^i \Gamma_{\mu v}^\lambda(p) A_j^v \\ &= \sum_{\lambda, v} y_\lambda^i(u) \Gamma_{\mu v}^\lambda(\pi(u)) x_j^v(u). \end{aligned}$$

This accounts for the coefficient of $d(x^\mu \circ \pi)$ in the desired expression for ω_j^i .

Now let us write

$$\sigma(E_\beta^\alpha)(u) = \sum_{\lambda, j} a_j^\lambda \frac{\partial}{\partial x_j^\lambda} \Big|_u$$

(clearly each $\partial/\partial x_j^\lambda$ is vertical, since $\pi_*(\partial/\partial x_j^\lambda)(f) = \partial(f \circ \pi)/\partial x_j^\lambda = 0$, as $f \circ \pi$ is constant on fibres; hence the $\partial/\partial x_j^\lambda$ span V_u). Then

$$\begin{aligned} a_j^\lambda &= \sigma(E_\beta^\alpha)(x_j^\lambda)(u) \\ &= \sigma_{u*}(E_\beta^\alpha)(x_j^\lambda) \\ &= \lim_{h \rightarrow 0} \frac{x_j^\lambda(u \cdot \exp h E_\beta^\alpha) - x_j^\lambda(u)}{h} \\ &= x_j^\lambda(u \cdot E_\beta^\alpha) = \lambda^{\text{th}} \text{ component with respect to } s(\pi(u)) \text{ of } (u \cdot E_\beta^\alpha)_j \\ &= \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \sum_\gamma u_\gamma (E_\beta^\alpha)_j^\gamma \\ &= \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad \text{''} \quad u_\beta \delta_j^\alpha \\ &= x_\beta^\lambda \delta_j^\alpha. \end{aligned}$$

So

$$\sigma(E_\beta^\alpha)(u) = \sum_\lambda x_\beta^\lambda \frac{\partial}{\partial x_\alpha^\lambda} \Big|_u.$$

Hence

$$\begin{aligned}
 \delta_\beta^i \delta_j^\alpha &= (E_\beta^\alpha)_j^i = \omega_j^i(\sigma(E_\beta^\alpha)(u)) \\
 &= \omega_j^i \left(\sum_\lambda x_\beta^\lambda \frac{\partial}{\partial x_\alpha^\lambda} \Big|_u \right) \\
 &= \sum_\lambda x_\beta^\lambda \omega_j^i \left(\frac{\partial}{\partial x_\alpha^\lambda} \Big|_u \right).
 \end{aligned}$$

This is easily seen to account for the other term, $\sum_\lambda y_\lambda^i dx_j^\lambda$, in the expression for ω_j^i . ♦

ADDENDUM 1

THE TANGENT BUNDLE
OF THE BUNDLE OF FRAMES

The existence of a connection on the bundle of frames $F(M)$ has some interesting implications.

23. PROPOSITION. If there is a connection on $F(M)$, then the tangent bundle of $F(M)$ is trivial.

PROOF. The $n^2 + n$ vector fields $\sigma(E_j^i)$ and $B(e_i)$ are everywhere linearly independent. ♦

24. COROLLARY. If there is a connection on $F(M)$, then M is paracompact.

PROOF. Since the tangent bundle of $F(M)$ is trivial, there is clearly a Riemannian metric on $F(M)$. So $F(M)$ is metrizable (Theorem I.9-7), so every component of $F(M)$ is σ -compact (Theorem I. A-1). Since $\pi: F(M) \rightarrow M$ is a continuous map onto M , it follows that each component of M is σ -compact, so M is metrizable (Theorem I. A-1). ♦

Of course, a non-paracompact manifold M may still have a connection in *some* principal bundle $\pi: P \rightarrow M$ with group G . For example, the trivial bundle $M \times G \rightarrow M$ has an obvious connection (the horizontal vectors Y are those with $\pi_{2*}Y = 0$, where $\pi_2: M \times G \rightarrow G$ is projection on the second coordinate).

If M is paracompact, then there is a connection in any principal bundle $\pi: P \rightarrow M$; this can be proved using partitions of unity, noting that any convex combination of Ehresmann connections is also a connection (a convex combination of connections ω_i makes sense, since the values of ω_i are in the vector space \mathfrak{g}).

For the special case of a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, we have a specific connection in $F(M)$, the Levi-Civita connection. It is easy to see that the Levi-Civita connection for a Riemannian metric can also be defined for an indefinite Riemannian metric (a non-degenerate inner product in each tangent space); it is the unique symmetric connection for which parallel translation is an isometry, and the Γ_{ij}^k are given by exactly the same formula as the Christoffel symbols. So Corollary 24 implies

25. COROLLARY. If M has an indefinite Riemannian metric, then M is paracompact.

This result cannot be proved by modifying the proof of Theorem I.9-7, for there may be paths of negative length between two points. For another proof of the result, see Sachs and Wu, *General Relativity for Mathematicians*, section 8.2.1.

ADDENDUM 2

COMPLETE CONNECTIONS

If ω is a connection on $F(M)$, we define a **geodesic**, as usual, to be a curve c such that dc/dt is parallel along c . The connection ω is **complete** if every geodesic segment can be extended to \mathbb{R} . Unlike the Levi-Civita connection for a Riemannian metric, a general connection on $F(M)$ may not be complete even if M is compact. To construct an example, we begin with the Levi-Civita connection for the metric $\langle \cdot, \cdot \rangle = e^x dx \otimes dx$ on \mathbb{R} . A geodesic c has constant squared length

$$\langle c'(t), c'(t) \rangle_{c(t)} = e^{c(t)} c'(t)^2,$$

so $e^{c(t)/2} c'(t)$ must be constant, and thus

$$c(t) = 2 \log(a + bt).$$

Thus, $\langle \cdot, \cdot \rangle$ is not complete (the geodesics run off to infinity in a finite amount of time). If we identify the bundle of frames $F(\mathbb{R})$ with $\mathbb{R} \times (\mathbb{R} - \{0\})$, then the integral manifolds of the horizontal distribution H of our connection are the sets

$$\begin{aligned} \{(c(t), c'(t))\} &= \left\{ \left(2 \log(a + bt), \frac{2b}{a+bt} \right) \right\} \\ &= \{(x, 2be^{-x/2})\}. \end{aligned}$$

Now we will identify x with $x + 1$ for all $x \in \mathbb{R}$; the resulting manifold is a circle S^1 and we have a map $\pi: \mathbb{R} \rightarrow S^1$ given by $\pi(x) =$ equivalence class of x . The distribution H on $F(\mathbb{R})$ gives rise to an obvious distribution on $F(S^1)$, which determines a connection on $F(S^1)$. The geodesics for this connection are the curves

$$\pi \circ c(t) = \pi(2 \log(a + bt)).$$

They cannot be extended to all of \mathbb{R} , since they go around S^1 infinitely often in a finite amount of time.

There are other anomalies for general connections. For example, even though a connection ω is complete, it may not be possible to join every pair of points with a geodesic. If we consider a Lie group G , it is easy to see that there is a unique connection ω on $F(G)$ which makes all left invariant vector fields parallel. Then geodesics through e are just 1-parameter subgroups. The connection is complete, since 1-parameter subgroups can be extended to all of \mathbb{R} , but (Problem I.10-27) it is not necessarily true that every element lies on a 1-parameter

subgroup. It is not known whether every two points can be joined by a geodesic in a compact manifold with a complete connection.

Finally, it should be pointed out that results about geodesics for the Levi-Civita connection of a Riemannian metric need not hold for the Levi-Civita connection of an indefinite Riemannian metric. For example, there is an indefinite metric whose Levi-Civita connection is complete but for which not every pair of points can be joined by a geodesic (see J. W. Smith, *Lorentz structures on the plane*, Trans. Amer. Math. Soc. **95** (1960), 226–237).

ADDENDUM 3

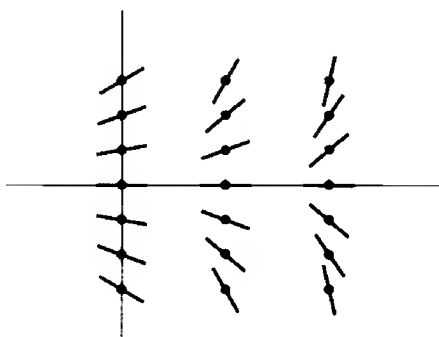
CONNECTIONS IN VECTOR BUNDLES

Suppose $\pi: E \rightarrow M$ is a C^∞ vector bundle over M . We have already seen how to form a principal bundle $\varpi: F(E) \rightarrow M$, in which* every $u \in \varpi^{-1}(p)$ is a frame for $\pi^{-1}(p)$. We can then define a **connection** in the original vector bundle E to be a connection in the principal bundle $F(E)$. However, there is a more direct way of defining a connection in E , which makes certain things work out more simply, since the bundle of frames contains a lot of superfluous stuff in it.

For a vector bundle $\pi: E \rightarrow M$ (as for a principal bundle) we define the **vertical subspace** $V_e \subset E_e$ at any point $e \in E$ to be the image $i_*(\pi^{-1}(p)_e)$, where $p = \pi(e)$; vectors in V_e are called **vertical**, and V_e is clearly the kernel of $\pi_*: E_e \rightarrow M_p$. For every frame $u \in \varpi^{-1}(p)$ and every $\xi \in \mathbb{R}^n$ we also have $u(\xi) \in \pi^{-1}(p)$ defined by $u(\xi) = \sum_i \xi^i \cdot u^i$. Finally, for every non-zero $\alpha \in \mathbb{R}$, we let $\bar{\alpha}: E \rightarrow E$ be the diffeomorphism defined by $\bar{\alpha}(e) = \alpha \cdot e$. We define a **connection** in E to be a distribution H such that

- (1) $E_e = H_e \oplus V_e$ for all $e \in E$
- (2) $\bar{\alpha}_*(H_e) = H_{\bar{\alpha}(e)} = H_{\alpha \cdot e}$ for all non-zero $\alpha \in \mathbb{R}$
- (3) H is a C^∞ distribution.

The subspace H_e is called the **horizontal subspace** at e , and vectors in H_e are called **horizontal**. Applying (2) with any $\alpha \neq 1$, and using a local trivialization, we see that if $s: M \rightarrow E$ is the zero section, then $H_{s(p)}$ must be $s_*(M_p)$.



We will first show how such a connection arises from a connection ω on the principal bundle $F(M)$. Let H be the distribution on $F(M)$ determined by ω .

*This bundle is a special case of the “associated principal bundle”, which is used in the theory of fibre bundles.

Given $e \in \pi^{-1}(p)$, choose any frame $u \in \pi^{-1}(p)$; then there is a unique $\xi \in \mathbb{R}^n$ with $u(\xi) = e$. Now we can define a map $\phi_\xi: F(E) \rightarrow E$ by $\phi_\xi(v) = v(\xi)$ for all $v \in F(E)$; this map takes fibres of $F(E)$ to fibres of E . We let $H_e = \phi_{\xi*}(H_u)$. This is well-defined, for if we choose $u \cdot A$ instead of u , then we must choose $A^{-1} \cdot \xi$ instead of ξ ; since

$$\phi_{A^{-1} \cdot \xi}(v) = v(A^{-1} \cdot \xi) = v \cdot A^{-1}(\xi),$$

we have $\phi_{A^{-1} \cdot \xi} = \phi_\xi \circ R_{A^{-1}}$, so

$$\begin{aligned} (\phi_{A^{-1} \cdot \xi})_*(H_{u \cdot A}) &= \phi_{\xi*} R_{A^{-1}*}(H_{u \cdot A}) \\ &= \phi_{\xi*} H_u, \quad \text{by Proposition 4.} \end{aligned}$$

We leave it to the reader to verify that these well-defined H_e do satisfy conditions (1), (2), (3).

We also want to show that every connection H in E does arise in this way from some connection H in $F(E)$. To do this, we first consider parallel translation. If $c: [0, 1] \rightarrow M$ is a curve, then the parallel translation

$$\tau_t: \pi^{-1}(c(0)) \rightarrow \pi^{-1}(c(t))$$

determined by H is defined in the obvious way: $\tau_t(e) = c^\dagger(t)$ where $c^\dagger: [0, 1] \rightarrow E$ is the unique curve with $c^\dagger(0) = e$ such that each $c^{\dagger'}(t)$ is horizontal; the existence of c^\dagger in E is proved similarly to the existence of c^* in a principal bundle. Clearly τ_t is a diffeomorphism, whose inverse is the parallel translation along the reversed part of c from $c(t)$ to $c(0)$. From condition (2) for H it follows that $\tau_t(\alpha \cdot e) = \alpha \cdot \tau_t(e)$ for $\alpha \neq 0$. This is true even for $\alpha = 0$, i.e., τ_t takes the zero vector at $c(0)$ to the zero vector at $c(t)$; this follows from the fact that $H_{s(p)} = s_*(M_p)$ when s is the zero section. It now follows that τ_t is a *vector space isomorphism*, because of the following

CLEVER OBSERVATION: If $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at 0 and satisfies $f(\alpha \cdot v) = \alpha f(v)$ for all $v \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, then f is linear.

PROOF. Let $T = Df(0): \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then

$$\begin{aligned} T(v) &= \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} [f(\alpha v) - f(0)] = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} f(\alpha v) \\ &= \lim_{\alpha \rightarrow 0} f(v) = f(v). \quad \spadesuit \end{aligned}$$

We now define the connection H in $F(E)$. Let $u \in \varpi^{-1}(p)$. For every C^∞ curve $c: [0, 1] \rightarrow M$ with $c(0) = p$, we define the curve $c^*: [0, 1] \rightarrow F(E)$ by

$$(c^*(t))_i = c^\dagger(t)(u_i) \quad i = 1, \dots, n,$$

the subscript i denoting the i^{th} component of the frame. Then we define H_u to be the set of all vectors $c^*(0)$ for all such curves c . The reader may verify that H is an Ehresmann connection, and that it gives rise to the connection H on E (the fact that τ_t is a vector space isomorphism is used to prove that $R_{A*}H_u = H_{u \cdot A}$).

Given a connection H on E , we use the direct sum decomposition $E_e = H_e \oplus V_e$, to define the **horizontal component** hX and **vertical component** vX of any tangent vector $X \in E_e$. Condition (2) for H is equivalent to the equation $\bar{\alpha}_* hX = h(\bar{\alpha}_* X)$, or to $\bar{\alpha}_* vX = v(\bar{\alpha}_* X)$. Remember that the tangent space $\pi^{-1}(p)_e$ at e of the vector space $\pi^{-1}(p)$ can be identified with $\pi^{-1}(p)$ itself. Since $vX \in i_*(\pi^{-1}(p)_e)$, we therefore can, and henceforth will, regard vX as an element of $\pi^{-1}(p)$.

The equation $\bar{\alpha}_* vX = v(\bar{\alpha}_* X)$ then becomes

$$(*) \quad v(\bar{\alpha}_* X) = \alpha \cdot vX.$$

Now consider a section $s: M \rightarrow E$ of the bundle $\pi: E \rightarrow M$, and a vector $X_p \in M_p$. We define

$$\nabla_{X_p} s = v s_*(X_p) \in \pi^{-1}(p).$$

It is clear that $\nabla_{X_p} s$ is linear in X_p and in s . For a C^∞ function $f: M \rightarrow \mathbb{R}$, the analogue of Proposition 3 is the formula

$$(**) \quad (f \cdot s)_*(X_p) = \overline{f(p)}_*(s_* X_p) + X_p(f) \cdot s(p),$$

where $s(p) \in \pi^{-1}(p)$ is identified with a tangent vector in $\pi^{-1}(p)$ at $f(p) \cdot s(p)$; to prove this formula, we introduce a local trivialization, and observe that it becomes the product rule for the derivative. If we take the vertical component of both sides of (**) and use (*), we find that

$$\nabla_{X_p}(f \cdot s) = f(p) \nabla_{X_p} s + X_p(f) \cdot s(p).$$

Thus ∇ is a Koszul connection. If the connection H in E comes from the connection H in $F(E)$, then this ∇ is the same as that determined by H ; this is an easy consequence of Lemma 8 (although Lemma 8 is concerned only with $F(M) = F(TM)$, it is easy to see that it generalizes to any $F(E)$).

Finally, we point out that if we are given ∇ , then H_e is simply $\{s_*(X_p) : p = \pi(e) \text{ and } \nabla_{X_p} s = 0\}$; it can also be defined as $s_*(M_p)$, where s is a section such that $\nabla s(p) = 0$.

ADDENDUM 4

FLAT CONNECTIONS

Consider the trivial principal bundle $\pi: M \times G \rightarrow M$, and let $\pi_2: M \times G \rightarrow G$ be projection on the second factor. We can define the **canonical flat connection** H on $M \times G$ by letting $H_{(p,a)} = \ker \pi_{2*}: (M \times G)_{(p,a)} \rightarrow G_a$. It is easy to see that the corresponding \mathfrak{g} -valued 1-form ω on $M \times G$ is given by $\omega = \pi_2^*(\omega')$, where ω' is the natural \mathfrak{g} -valued 1-form on G (defined on pg. I.403). Using the equations of structure of G , in the form given on pg. I.404, we now have, for all $Y_1, Y_2 \in (M \times G)_u$

$$\begin{aligned} d\omega(Y_1, Y_2) &= d\pi_2^*(\omega')(Y_1, Y_2) \\ &= \pi_2^*(d\omega')(Y_1, Y_2) \\ &= d\omega'(\pi_{2*}Y_1, \pi_{2*}Y_2) \\ &= -[\omega'(\pi_{2*}Y_1), \omega'(\pi_{2*}Y_2)] \\ &= -[\omega(Y_1), \omega(Y_2)]. \end{aligned}$$

Comparing with the structural equations of the principal bundle $M \times G$ (Theorem 16), we see that for this connection ω we have $\Omega = 0$.

Conversely, suppose $\pi: P \rightarrow M$ is a principal bundle with group G , and a connection ω such that $\Omega = 0$. By Proposition 18, the distribution H corresponding to ω is integrable. From this it is easy to see that around any point $p \in M$ there is a neighborhood U and a diffeomorphism $t: \pi^{-1}(U) \rightarrow U \times G$ of the form $t(u) = (\pi(u), \phi(u))$, where $\phi(u \cdot a) = \phi(u) \cdot a$, such that $t_*(H_u)$ is the horizontal subspace at $t(u)$ for the canonical flat connection in $U \times G$. (One part of our final proof of the Test Case essentially used this fact.)

It should be noted that the second structural equation of Euclidean space (Proposition 7-1), with which we began our whole investigation of moving frames, could have been deduced from the equations of structure of $GL(n, \mathbb{R})$, by a process essentially equivalent to the deduction just given for the equation $d\omega(Y_1, Y_2) = -[\omega(Y_1), \omega(Y_2)]$. To obtain the *first* structural equation, we would have had to consider the equations of structure for the group of affine motions [an “affine motion” is a translation followed by an element of $GL(n, \mathbb{R})$]. In general, for a connection ω on $F(M)$, the torsion form and the first structural equation for ω can be interpreted in terms of a connection in the bundle $A(M)$ of affine frames of M , where an “affine frame” of M_p is a pair (v, u_1, \dots, u_n) , for $v \in M_p$ and (u_1, \dots, u_n) a frame for M_p . For this interpretation, the reader is referred to Kobayashi and Nomizu, *Foundations of Differential Geometry*, Vol. 1, pp. 125–130.

NOTATION INDEX

CHAPTER 1

b	28
dP	37
k_1, \dots, k_{n-1}	45
n	6, 27
P	37
P^{-1}	37
\overline{pq}	13
$\mathrm{SL}(n, \mathbb{R})$	38
s	1
$\mathfrak{sl}(n, \mathbb{R})$	38
t	6
$v \times w$	29
$\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$	34
α_c	46
$\gamma'(s)$	1
κ	6, 27
κ	41, 47
μ	16
σ	39, 46
τ	28
ω	36

CHAPTER 2

c_X	49
c_ϕ	53
κ_X	49
κ_ϕ	53
ν	53

CHAPTER 3A

$\cos(1)L$	62
dx	64
$d\delta_X$	94
$d\sigma$	74
L	62
v_A	62
X, Y, Z	62
δ_X	94

ν	68
ω	102
(1), (2), (3)	56

CHAPTER 3B

dV	116
$d\nu$	121
E, F, G	127
G	136
g	138
I	122
II	122
$K(p)$	114, 116
l, m, n	128
(r, φ)	136
ν	112
$\theta(s)$	137
(ρ, ϕ)	136
σ'	116
$\mathbb{I} \quad \mathbb{I}$	139

CHAPTER 4A

$A(X, Y)$	170
$Q(v_p, w_p)$	169
$\ v_p, w_p\ $	169

CHAPTER 4D

$K(W)$	194
$R(Y, Z)X$	189
R_{ijkl}	190
R^i_{jkl}	188

CHAPTER 4. ADDENDUM

\mathcal{F}	202
$f_{**}(v)$	201
g_{ij}	202
g^{ij}	204
$\Gamma^k_{ij}, [ij, k]$	204

CHAPTER 5

$A_{i_1 \dots i_k; h}^{j_1 \dots j_l}$	210
$A_{i_1 \dots i_k; h\eta}^{j_1 \dots j_l}$	213
\exp	223
$f; h$	211
T_{ij}^k	221
Γ_{ij}^k	221
$\lambda_{i; h}$	210
λ_{ih}^j	211
$\lambda_{i, h}^j$	211
$\lambda_{i h}^j$	211
$\lambda^j_{; h}$	210

CHAPTER 6

$A(X, Y)$	249
$D(X, Y)$	249
$\frac{DV}{dt}$	232
$\frac{DV}{\partial x}$	238
$R(X, Y)Z$	239
$S(X, Y)$	249
$\ni A(X, Y, Z)$	243
$T(X, Y)$	236
Γ_{ij}^k	228
τ_t	234
$\frac{\partial s}{\partial x}$	238
$\frac{\partial s}{\partial y}$	238
∇	227
∇A	230
∇Y	228
$\nabla_X A$	229
$\nabla_X Y$	227

$\nabla_{X_p} A$	235
------------------	-----

$\nabla_{X_p} Y$	227
------------------	-----

CHAPTER 7

(A_j^i)	261
$A \cdot v$	262
dP	260, 264
G	279
g	279
$K(W)$	277
$L(X, V)$	277
$P: \mathbb{R}^n \rightarrow \mathbb{R}^n$	260
$R(X, Y)Z$	285
\mathbf{R}_{kij}^l	266
r	294
\mathbf{s}	290
$T(X, Y)$	285
\mathbf{T}_{jk}^i	287
(t, t^1, \dots, t^n)	270
$\mathbf{v} \cdot A$	262
$\mathbf{X} \cdot A$	262
Γ_{ij}^k	266
Θ^i	283
θ^i	260, 281
$\bar{\theta}^i$	271
$\bar{\bar{\theta}}^i$	276
ρ	293
Φ	270
Ω_j^i	266, 283
ω_j^i	260
$\bar{\omega}_j^i$	271
$\omega \wedge \eta$	262
$\omega \wedge \theta$	262
$\frac{\partial \bar{\theta}^i}{\partial t}, \frac{\partial \bar{\omega}_j^i}{\partial t}$	272
∇Y	285
$\angle(v, w)$	296
$\langle \ , \ \rangle$	276

CHAPTER 8

$\text{Ad}(a)$	309	$u: \mathbb{R}^n \rightarrow M_{\pi(u)}$	319
$A(M)$	349	V_e	346
$B(\xi)$	325	V_u	307
c_p	309	$v(Y)$	317, 348
c^*	318, 347	X^*	317
c^\dagger	347	$x_j^i(u)$	306
$D\alpha$	324	$x_\#$	306
E_i^j	322	$(y_j^i(u))$	338
f_Y	320	$\tilde{\alpha}$	346
$F(E)$	307	Θ	325
$F(M)$	306	Θ^i	328
H	315, 346	θ	324
H_e	346	θ^i	328
H_u	315	$\varpi: F(E) \rightarrow M$	307
$h(Y)$	317, 348	$\sigma(X)$	309, 311
$O(E)$	308	τ_t	320, 347
R_a	309	ϕ_ξ	347
$R(X_1, X_2)X_3$	330	Ω	324
$s \cdot a$	312	Ω_j^i	328
$\text{SF}(E)$	308	ω	315
$\text{SO}(E)$	308	ω_s	311
$T(X_1, X_2)$	329	(ω_j^i)	322, 328
$u \cdot A$	306	$\nabla_{X_p} s$	348
		$\nabla_{X_p} Y$	320

INDEX

- Absolute derivative, 211
- Absolute Differential Calculus, 209
- Acceleration vector, 247
- Act
 - effectively, 309
 - on a manifold on the right, 309
 - on the left, 262
 - on the right, 262
 - without fixed point, 309
- Action of the group of a principal bundle, 306
- Adapted moving frame, 270
- Affine
 - arclength, 39; *see also* Special affine arclength
 - curvature, 41; *see also* Special affine curvature
 - frame, 349
 - motion, 39, 349
 - special, 39
- Ambrose, W., 278, 329
- Analytic
 - curve, 46
 - surface, 297
- Angle, 296
 - rotation through an angle of θ , 50
- Arclength function, 1
- Arclength, special affine, 46
- Area of $v(A)$, 114
- Associated principal bundle, 346

- Ball, unit, 203
- Banach space, 14
 - norm, 14, 203
- Base space of principal bundle, 306
- Basic vector field, 325
- Bell, E. T., 149
- Bending invariant, 133
- Bertrand, J. and Puiseux, Formula of, 147
- Bianchi's
 - first identity, 224, 244, 288, 334
 - identity, 225
 - second identity, 224, 244, 288, 334
- Binormal, 28
- Bolyai, J., 163
- Bundle
 - associated principal, 346
 - fibre, 346
 - of frames, 306
 - tangent bundle of, 342
 - principal, 306
 - connection in, 315
 - vector, connection in, 346

- Calculus of Tensors, 209
- Canonical
 - flat connection, 349
 - form, 324
- Cartan, Élie, 259, 302, 304
- Cartan connection, 281, 311
- Chevalley, C., 144
- Christoffel, E. B., 186, 209
- Christoffel symbols, 186
- Circle
 - geodesic, 147
 - osculating, 6, 27
- Classical connection, 221
- Clever observation, 347
- Clifford, W. K., 150
- Closed curve, 12
 - convex, 12
 - parallel translation of a vector along, 243
 - simple, 12
- Compatible connection, 236
- Complete
 - connection, 344
 - set of invariants, 39
- Component
 - horizontal, 317, 348
 - vertical, 317, 348
- Conformal, 296
- Conformally equivalent, 297
- Conic
 - hyperosculating, 43
 - section, 43

Connection

- canonical flat, 349
- Cartan, 281, 311
- classical, 221
- compatible with metric, 236
- complete, 344
- difference tensor of two, 249
- Ehresmann, 315
 - convex combinations of, 342
- flat, 349
 - canonical, 349
- form, 317
- forms, 266, 281
- in principal bundle, 315
 - over paracompact space, 342
- in vector bundle, 346
- Koszul, 227
- Levi-Civita, 238, 288
- projectively equivalent, 250
- reason for terminology, 234
- summary of different definitions of, 337
- symmetric, 222
- with the same geodesics, 249

Constant

- curvature, 10, 290
- special affine curvature, 41

Convex

- combination of Ehresmann connections, 342
- curve, 12
- function, 203
- set, 13
 - in \mathbb{R}^3 , 52

Courant, R., 126

Covariant

- derivative, 211, 223
 - of a vector field along a curve, 232
- differential, 324
- differentiation, rules for, 212

Curvature

- constant, 10, 290
- curve of double, 28
- determines the metric, 277
 - two-dimensional case, 279
- double, 28

first, 28

form, 324

forms, 266, 283

functions

- of a curve in \mathbb{R}^n , 45

- special affine, 47

Gaussian, 116

- formulas for, 119, 120, 126, 127, 129

- in geodesic polar coordinates, 138

mean, 133

of a cylinder, 116

of a plane, 115

of a plane curve, 6

- first, 28

- formulas for, 7, 8

- global, 16

- second, 28

- total, 18

of a space curve, 27

of a sphere, 115

of a surface, 70, 114, 116, 157

- total, 68

- of the intersection of a plane and a surface, 49

- Riemann's invariant definition of, 254

second, 28

sectional, 194

- preserving diffeomorphism, 279

tensor, 189, 223, 239, 286, 330

total, 18, 68

Curve

analytic, 46

closed, 12

- parallel translation of a vector along, 243

convex, 12

curvature of, 6, 27

- first, 28

- global, 16

- second, 28

- total, 18

direction of, 2

global results about, 12 ff.

going in opposite direction, 7

horizontal, 318

- Curve (*continued*)
 - immersion of, 1
 - in plane, 1
 - in space, 24
 - inside of, 15
 - length of, 200
 - lift of, 318
 - of double curvature, 28
 - parallel translation of fibres along, 319
 - parallel translation of vector along, 234
 - parallel vector field along, 233
 - projections of, 30
 - rotation index of, 19
 - simple, 12
 - Taylor expansion of, 27, 31
 - vector field along, 231
 - covariant derivative of, 232
 - C^∞ , 231
 - vertex of, 23
- Cylinder, curvature of, 116
- C^∞ Minkowski metric, 200
- C^∞ vector field along curve, 231
- Debauch of indices, 209 ff.
- Dedekind, R., 149, 150
- Derivative
 - absolute, 211
 - covariant, 211
 - “partial”, 272
- Determinant, 143
- Development of a surface, 90
- Difference tensor of two connections, 249
- Differentiable surface, 62
- Differential, covariant, 324
- Diquet, 147
- Direction of a curve, 2
- Dogma, 143
- Double curvature, 28
- Dual
 - form, 324
 - forms, 260, 264, 281
- Effectively, act, 309
- Ehresmann, C., 305
- Ehresmann connection, 315
 - convex combinations of, 342
- Eigenvalues, 125
 - minimax definition of, 126
- Eigenvectors, 125
- Ellipse, 23
- Ellipsoid, 205
 - of smallest volume, 206
- Equations of structure of Lie group, 349
- Euclidean motion, proper, 11, 34, 43
- Euclidean space
 - structural equations of, 261
 - as integrability conditions, 262
- Euler, L., 49
- Euler’s Theorem, 50, 121, 123, 125, 126
- Exterior algebra, 144
- Fancy free, 124
- Fibre, 307
 - bundle, 346
 - parallel translation of along a curve, 319
- Finsler, P., 165
- Finsler metric, 165, 202
- First curvature, 28
- First fundamental form, 122
- First structural equation, 327
- First structural equation of Euclidean space, 349
- First Variation Formula, 247
- Fixed point, act without, 309
- Flat
 - connection, 349
 - manifold, 179, 184
 - torus, 179
- Four Vertex Theorem, 23
- Frame, 259, 306
 - affine, 349
 - bundle of, 306
 - moving, 259

- Frenet, F., 34; *see also* Serret-Frenet formulas
- Fundamental form, *see* First and Second fundamental form
- Fundamental vector field, 311

- Gauss, C. F., 49, 55 ff., 149, 150, 297
- Gauss map, 112
- Gauss' Lemma, 247
- Gaussian curvature, 116
 - formulas for, 119, 126, 127, 129
- Geodesic, 135, 223, 246
 - circle, 147
 - connections with the same, 249
 - on a surface in \mathbb{R}^3 , 135
 - polar coordinates, 290
 - curvature in, 138
 - triangle, 141
- Geometry, non-Euclidean, 163
- Global formulation of the curvature function, 16
- Global results about curves, 12 ff.
- Gram-Schmidt orthonormalization, 43, 260
- Great circles, triangle of, 134

- Hahn-Banach Theorem, 14
- Hairy calculation, 174
- Helix, 32
 - "left handed", 33
 - "right handed", 33
 - rotation of, 33
- Hessian, 201
- Horizontal
 - component, 317, 348
 - curve, 318
 - subspace, 315, 346
 - vector, 315, 346
- Huygens, C., 9
- Hyperosculating conic, 43

- Immersion, 1
- Indefinite Riemannian metric, Levi-Civita connection for, 342
- Index, rotation, 19
- Inequality
 - Schwarz, 135
 - triangle, 203
- Infinitesimal triangle, 74, 100
- Inside of a curve, 15
- Integrability conditions, 262
- Invariance of curvature under proper Euclidean motions, 39
- Invariants, complete set of, 39
- Inward pointing normal, 52
- Isothermal coordinates, 297
- Isotropic, 291

- Jacobi identity, 245

- Kobayashi, S., 349
- Kobayashi and Nomizu, 349
- Koszul, J. L., 227
- Koszul connection, 227
- Kulkarni, R. S., 279

- Left handed helix, 33
- Left, act on, 262
- Leibniz, G. W., 9
- Length of curve, 200
- Levi-Civita, T., 209, 234
- Levi-Civita connection, 238
 - for an indefinite Riemannian metric, 342
- Lie group
 - equations of structure of, 349
 - review of, 36 ff.
- Lift
 - of curve, 318
 - of vector field, 317
- Linear group, special, 38
- Lobachevsky, N. I., 163
- Local triviality, 307

- Manifold
 - flat, 179, 184
 - Riemannian, 165
- Matrix notation, modified, 261
- Mean curvature, 133
- Metric
 - determined by the curvature, 277 ff.
 - Finsler, 165, 202
 - Levi-Civita connection for indefinite Riemannian, 342
 - Minkowski, 200
- Meusnier, J. B., 52
- Meusnier's Theorem, 54, 123
- Milnor, J. W., 195
- Minimax definition of eigenvalues, 126
- Minkowski metric, 200
- Modified matrix notation, 261
- Motion
 - affine, 349
 - Euclidean, 11, 34, 43
 - special affine, 39
- Moving frame, 259
 - adapted, 270
 - natural, 259
 - orthonormal, 260, 264, 266
- Natural moving frame, 259
- Natural parameter
 - for curves under the group of Euclidean motions, 38
 - for curves under the group of special affine motions, 40
- Natural g -valued 1-form, 36
- Newton, I., 9
- Nomizu, K., 349; *see also* Kobayashi and Nomizu
- Non-Euclidean geometry, 163, 301
- Norm
 - Banach space, 203
- Normal
 - coordinates, 302; *see also* Riemannian normal coordinates
 - map, 112
 - plane, 30
- vector
 - inward pointing, 52
 - unit, 62
- Opposite direction, curve going in, 7
- Orthonormal moving frame, 260, 264, 266
- Orthonormalization, Gram-Schmidt, 43, 260
- Osculate, 6
- Osculating
 - circle, 6, 27
 - parabola, 43
 - plane, 25
- Parabola, osculating, 43
- Parallel
 - translation of fibre along curve, 319
 - translation of vector along closed curve, 243
 - translation of vector along curve, 234
 - vector field, 243
 - vector field along curve, 234
- Parallelepiped, volume of, 56
- Parameter
 - natural for curves under group of Euclidean motions, 38
 - natural for curves under group of special affine motions
 - parameterized surface, 238
 - vector field along, 238
- "Partial derivatives", 272
- Permutation, sign of, 144
- Plane
 - curvature of, 115
 - normal, 30
 - osculating, 25
 - rectifying, 30
- Point, 37

- Polar coordinates
 - geodesic, 290
 - curvature in, 138
 - structural equations in, 272
- Polarization, 185
- Principal axes, 205
- Principal bundle, 306
 - action of group of, 306
 - associated, 346
 - base space of, 306
 - connection in, 315
 - local triviality of, 307
 - over paracompact spaces, connections in, 342
 - projection map of, 306
 - total space of, 306
 - trivial, 307
- Principal normal, 28
- Projection map of a principal bundle, 306
- Projections of a curve, 30
- Projectively equivalent connections, 250
- Proper Euclidean motion, 11
- Puiseux, V. A. (Formula of Bertrand and Puiseux), 147
- Pythagorean Theorem, 204

- Rectifying plane, 30
- Repère mobile, 259
 - fundamental principle of, 270
- Ricci, G., 209
- Ricci Calculus, 209 ff.
- Ricci's identities, 214, 239
- Ricci's lemma, 213
- Riemann, G. F. B., 149 ff., 209, 295, 301, 302
- Riemann curvature tensor, 189
- Riemannian manifold, 165
 - locally isometric, 165
 - structural equations of, 267
- Riemannian metric, 165
 - connection compatible with, 236
 - determined by the curvature, 277 ff.
 - indefinite, 342
 - Levi-Civita connection for, 342
- Riemannian normal coordinates, 166
- Right
 - act on, 262
 - act on M on, 309
- Right handed helix, 33
- Rotation, 11, 34
 - index, 19
 - of a helix, 33
 - through an angle of θ , 50

- Scalar triple product, 29
- Schur, F., 291
- Schwarz inequality, 135
- Second curvature, 28
- Second fundamental form, 122
 - symmetry of, 123
- Second structural equation, 327
 - of Euclidean space, 349
- Section, 311
- Sectional curvature, 194
 - preserving diffeomorphism, 279
- Self-adjoint, 125
- Serret, J. A., 34
- Serret-Frenet formulas, 34
- Sign of a permutation, 144
- Simple curve, 12
- Singer, I. M., 329
- Smith, D. E., 150
- Smith, J. W., 345
- Space curve, curvature of, 27
- Special affine
 - arclength, 39, 46
 - curvature, 41
 - constant, 41
 - curvature functions, 47
 - motion, 39
- Special linear group, 38
- Spectral theorem, 125
- Sphere
 - curvature of, 115
 - unit, of Minkowski metric, 200
- Stokes' Theorem, 143

- Structural equation(s), 287
 - first, 327
 - in polar coordinates, 272
 - of Euclidean space, 261
 - first, 349
 - second, 349
 - of Riemannian manifold, 267
 - second, 327
- Subgroups of $GL(n, \mathbb{R})$, ω for, 37
- Subspace
 - horizontal, 315, 346
 - vertical, 307, 346
- Summary of different definitions of
 - connections, 337
- Support line, 13
- Surface
 - curvature determines metric, 279
 - geodesic on, 135
 - parameterized, 238
- Symmetric connection, 222, 238
- Symmetry of second fundamental
 - form, 123
- Tangent bundle of $F(M)$, 342
- Tangent space of $O(n)$, 35
- Taylor expansion of a curve, 27, 31
- Taylor polynomial approximations of
 - g_{ij} , 165
- Tensors, Calculus of, 209
- Test Case, 179
 - first version, 197
 - second version, 217
 - third version, 241
 - fourth version, 268
 - fifth version, 274
 - sixth version, 275
 - seventh version, 333
- Theorema Egregium, 132, 143
- Torsion, 28
 - form, 325
 - forms, 283
 - formula for, 29, 30
 - tensor, 221, 236, 286, 330
- Torus, flat, 179
- Total curvature, 18
- Total space of a principal bundle, 306
- Triangle
 - geodesic, 141
 - inequality, 203
 - infinitesimal, 74, 100
 - of great circles, 134
- Triple product, scalar, 29
- Trivial principal bundle, 307
- Two-dimensional manifolds; curvature
 - determines the metric, 279
- Two-parameter variation, 256
- Unit
 - ball, of Finsler metric, 203
 - normal vector, 62
 - sphere, of Minkowski metric, 200
- Variation
 - two-parameter, 256
 - vector field, 247
- Vector
 - horizontal, 315, 346
 - vertical, 307, 346
- Vector bundle, connection in, 346
- Vector field
 - along a curve, 231
 - covariant derivative of, 232
 - along a parameterized surface, 238
 - basic, 325
 - fundamental, 311
 - lift of, 317
- Velocity vector, 247
- Vertex of a curve, 23
- Vertical
 - component, 317, 348
 - subspace, 307, 346
 - vector, 307, 346
- Volume of a parallelepiped, 56
- Weber, H., 150, 170
- Weingarten
 - equations, 124
 - map, 122
- Weyl, H., 170, 251

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME THREE
Third Edition



MICHAEL SPIVAK

PUBLISH OR PERISH, INC.



Houston, Texas 1999

Publish or Perish, Inc.
www.mathpop.com

Copyright © 1970, 1979, 1999 by Michael Spivak
All Rights Reserved

Volume 1 ISBN 0-914098-70-5
Volume 2 ISBN 0-914098-71-3
Volume 3 ISBN 0-914098-72-1
Volume 4 ISBN 0-914098-73-X
Volume 5 ISBN 0-914098-74-8

Printed in the United States of America

PREFACE

These final three volumes are regarded as constituting a single volume, with Chapters 1 to 6 in Volume III, Chapters 7 to 9 in Volume IV, and Chapters 10 to 13 in Volume V. After finishing this multi-volume, I felt somewhat like a man who has tried to cleanse the Augean stables with a Johnny-Mop. Leafing through Mathematical Reviews for the past thirty years, and gazing at the dignified tomes which represent the glories of the classical era, one quickly senses that Differential Geometry is a field of overwhelming extent, beyond the comprehension of any mortal. I suppose such lucubrations ought to buoy up one's spirit with admiration for human achievement, but I must confess that they usually lead me instead to a state of brooding melancholy.

Although the strident word "comprehensive" still stands emblazoned in the title, the Bibliography, in Volume V, will begin to give some idea how much as has necessarily been left out. There are also mini-bibliographies in Volumes III and IV for the works explicitly cited there. Problems have been restricted practically to the absolute minimum, basically facts left to the reader as exercises.

As a glance at the table of contents will show, Volume III is essentially a course in classical surface theory, the only difference being that Chapter 1 prepares the ground for applying the intrinsic geometry of Riemannian manifolds, which was discussed in Volume II. Although much space is devoted to classical material, much of the generalized material in Chapter 7 would be almost incomprehensible without the prior treatment of surface theory. The only exception is the second half of Chapter 2 (from page 75 on), which can (and probably should) be omitted completely without loss of continuity. For those who don't care for the motivational twiddle-twaddle, an introduction to "modern" differential geometry can be extracted from Chapters 1, 7 (parts D and E), 8, and 13, with an assist from Chapter 5, and the first halves of Chapters 9 and 12.

References like Theorem 6-3 or 7-2, when quoted in Volumes III or IV, say, refer to Chapter 6 of Volume III, and Chapter 7 of Volume IV, respectively. References to results of Volumes I and II, or page numbers from any other volume, carry an additional Roman numeral, e.g., Theorem I.6-3, or pg. IV.167.

As acknowledged in the first edition, I am grateful to the Sonderforschungsbereich Theoretische Mathematik in Bonn, the University of California at Berkeley, and the following individuals (as well as many readers and others whom I may have inadvertently omitted): R. Bassein, R. Bishop, R. Böhme, D. Bourghelia, E. Calabi, B. Cenk, S. S. Chern, M. do Carmo, P. Eberlein, J. Ehrbacher, W. Fulton, R. Gardner, P. Gilkey, R. Greene, R. Gulliver, R. Hartshorne, S. Hildebrandt, M. Hirsch, F. Hirzebruch, G. Hochschild, H. Jacobowitz, J. Kazdan, W. Klingenberg, S. Kobayashi, R. Kulkarni, B. Lawson, R. Maltz, P. Melvin, J. Milnor, T. Milnor, J. Moore, C. Morrey, J. Nash, L. Nirenberg, K. Nomizu, B. O'Neill, R. Osserman, T. Ōtsuki, R. Palais, H. Pittie, J. Polking, E. Portnoy, M. Protter, R. Reilly, R. Risch, H. Royden, E. Ruh, P. Ryan, R. Sacksteder, S. Sasaki, D. Schaeffer, J. Sjögren, C. Snugg, E. Spanier, J. Stallings, E. Thomas, F. Warner, J. Weiner, A. Weinstein, J. White, J. Wolf, H. Wu, A. Vasquez, D. Zagier, W. Ziller.

TABLE OF CONTENTS

Although the chapters are not divided into sections,
the listing for each chapter gives some indication
which topics are treated, and on what pages.

CHAPTER 1. THE FUNDAMENTAL EQUATIONS FOR HYPERSURFACES

Covariant differentiation in a submanifold of a Riemannian manifold	1
The second fundamental form, the Gauss formulas, and Gauss' equation; Synge's inequality	4
The Weingarten equations and the Codazzi-Mainardi equations for hypersurfaces	7
The classical tensor analysis description	12
The moving frame description	16
Addendum. Auto-parallel and totally geodesic submanifolds . . .	22
Problems	28

CHAPTER 2. ELEMENTS OF THE THEORY OF SURFACES IN \mathbb{R}^3

The first and second fundamental forms	31
Classification of points on a surface; the osculating paraboloid and the Dubin indicatrix	35
Principal directions and curvatures, asymptotic directions, flat points and umbilics; all-umbilic surfaces	48
The classical Gauss formulas, Weingarten equations, Gauss equation, and Codazzi-Mainardi equations	51
Fundamental theorem of surface theory	56
The third fundamental form	62
Convex surfaces; Hadamard's theorem	64
The fundamental equations via moving frames	68
Review of Lie groups	71

Application of Lie groups to surface theory; the fundamental equations and the structural equations of $SO(3)$	72
Affine surface theory; the osculating paraboloids and the affine invariant conformal structure	75
The special affine first fundamental form	82
Quadratic and cubic forms; apolarity	91
The affine normal direction; the special affine normal	97
The special affine Gauss formulas and special affine second fundamental form	103
The Pick invariant; surfaces with Pick invariant 0	111
The special affine Weingarten formulas	121
The special affine Codazzi-Mainardi equations; the fundamental theorem of special affine surface theory	128
Problems	134

CHAPTER 3. A COMPENDIUM OF SURFACES

Basic calculations	135
The classical flat surfaces	141
Ruled surfaces	146
Quadric surfaces	151
Surfaces of revolution	156
Rotation surfaces of constant curvature	161
Minimal surfaces	167
Addendum. Envelopes of 1-parameter families of planes	176
Problems	182

CHAPTER 4. CURVES ON SURFACES

Normal and geodesic curvature	187
The Darboux frame; geodesic torsion	191
Laguerre's theorem	193
General properties of lines of curvature, asymptotic curves, and geodesics	195
The Beltrami-Eininger theorem	200
Lines of curvature and Dupin's theorem	203
Conformal maps of \mathbb{R}^3 ; Liouville's theorem	208
Geodesics and Clairaut's theorem	211

Addendum 1. Special parameter curves	217
Addendum 2. Singularities of line fields	218
Problems	224
CHAPTER 5. COMPLETE SURFACES OF CONSTANT CURVATURE	
Hilbert's lemma; complete surfaces of constant curvature $K > 0$. .	233
Analysis of flat surfaces; the classical classification	
of developable surfaces	235
Complete flat surfaces	244
Complete surfaces of constant curvature $K < 0$	247
CHAPTER 6. THE GAUSS-BONNET THEOREM AND RELATED TOPICS	
The connection form for an orthonormal moving frame on a surface; the change in angle under parallel translation	261
The integral of $K dA$ over a polygonal region	265
The Gauss-Bonnet theorem; consequences	270
Total absolute curvature of surfaces	278
Surfaces of minimal total absolute curvature	280
Total curvature of curves; Fenchel's theorem, and the Fary-Milnor theorem	286
Addendum 1. Compact surfaces with constant negative curvature .	292
Addendum 2. The degree of the normal map	299
Problems	303
MINI-BIBLIOGRAPHY FOR VOLUME III	305
NOTATION INDEX	307
INDEX	309

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME THREE

CHAPTER 1

THE FUNDAMENTAL EQUATIONS FOR HYPERSURFACES

In this chapter we are going to begin by considering a very general situation. Let $i: M^n \rightarrow N^m$ be an immersion of an n -dimensional manifold M into an m -dimensional manifold N ; it is customary to refer to N as the “ambient space” and to define $m - n$ to be the **codimension** of M in N . We will be interested in the case where N has a Riemannian metric $\langle \cdot, \cdot \rangle$, so that M can be given the induced Riemannian metric $i^*\langle \cdot, \cdot \rangle$. (This setup is often described a little differently: we can begin with two Riemannian manifolds $(N, \langle \cdot, \cdot \rangle)$ and $(M, \langle \cdot, \cdot \rangle)$, and consider isometric immersions of M in N , that is, immersions $i: M \rightarrow N$ which satisfy $i^*\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle$.)

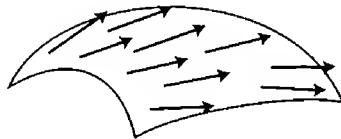
For every $p \in M$, we can consider the tangent space M_p as a subspace of the tangent space $N_{i(p)}$, by identifying M_p with $i_*M_p \subset N_{i(p)}$. Since all the results of this chapter are going to be local ones, it will simplify our notation considerably to assume that M is actually an imbedded submanifold of N , with $i: M \rightarrow N$ the inclusion map. We can then regard M_p as a subspace $M_p \subset N_p$. In the vector space N_p , with the inner product $\langle \cdot, \cdot \rangle_p$, the subspace M_p has an orthogonal complement $M_p^\perp \subset N_p$, and we can use the decomposition $N_p = M_p \oplus M_p^\perp$ to define two projections

$$\begin{aligned} \mathsf{T}: N_p &\rightarrow M_p && \text{(the tangential projection)} \\ \mathsf{\perp}: N_p &\rightarrow M_p^\perp && \text{(the normal, or perpendicular, projection)} \end{aligned}$$

with

$$X = \mathsf{T}X + \mathsf{\perp}X \quad \text{for all } X \in N_p.$$

Now let $X_p \in M_p$ be any vector, and let Y be a vector field on M which is everywhere “tangent to M ”, meaning that $Y_q \in M_q$ for $q \in M$. Then



$\nabla_{X_p} Y \in M_p$ is defined, where ∇ denotes the covariant differentiation in M which is determined by the induced Riemannian metric $i^*\langle \cdot, \cdot \rangle$. If ∇' denotes the covariant differentiation determined by $\langle \cdot, \cdot \rangle$ in the ambient space N , then $\nabla'_{X_p} Y$ is also well-defined; in fact, the value of $\nabla'_{X_p} Y$ depends only on the values of Y along some curve c with $c'(0) = X_p$. The relation between these two covariant differentiations is as nice as one could hope:

1. THEOREM. Let $i: M \rightarrow N$ be an immersion, where N has the Riemannian metric $\langle \cdot, \cdot \rangle$, and let ∇ and ∇' be the covariant derivatives for $(M, i^*\langle \cdot, \cdot \rangle)$ and $(N, \langle \cdot, \cdot \rangle)$, respectively. If $X_p \in M_p$, and Y is a vector field on M which is everywhere tangent to M , then

$$\nabla_{X_p} Y = \mathbf{T}(\nabla'_{X_p} Y).$$

PROOF. Let X, Y, Z be vector fields on N . Since ∇' is compatible with $\langle \cdot, \cdot \rangle$, we have (Corollary II.6-7)

$$\begin{aligned} X\langle Y, Z \rangle &= \langle \nabla'_X Y, Z \rangle + \langle Y, \nabla'_X Z \rangle \\ Y\langle Z, X \rangle &= \langle \nabla'_Y Z, X \rangle + \langle Z, \nabla'_Y X \rangle \\ -Z\langle X, Y \rangle &= -\langle \nabla'_Z X, Y \rangle - \langle X, \nabla'_Z Y \rangle. \end{aligned}$$

We also have $\nabla'_X Y - \nabla'_Y X = [X, Y]$, etc., and in particular, $\nabla'_X Y + \nabla'_Y X = 2\nabla'_X Y - [X, Y]$. Adding the above three equations, we thus obtain

$$\begin{aligned} (*) \quad X\langle Y, Z \rangle + Y\langle Z, X \rangle - Z\langle X, Y \rangle \\ = 2\langle \nabla'_X Y, Z \rangle - \langle [X, Y], Z \rangle + \langle [X, Z], Y \rangle + \langle [Y, Z], X \rangle. \end{aligned}$$

This equation shows that $\langle \nabla'_X Y, Z \rangle$ is completely determined by $\langle \cdot, \cdot \rangle$ (and is essentially equivalent to our proof of Lemma II.6-8).

Now consider three vector fields X, Y, Z on N which are tangent to M at all points of M , so that there are vector fields $\bar{X}, \bar{Y}, \bar{Z}$ with $i_*\bar{X}(p) = X(p)$, etc. On M we have the same equation (*) for the vector fields $\bar{X}, \bar{Y}, \bar{Z}$, but with $\nabla'_X Y$ replaced by $\nabla_{\bar{X}} \bar{Y}$. For a bracket term like $[\bar{X}, \bar{Y}]$ we have $[\bar{X}, \bar{Y}](p) = [X, Y](p)$ for $p \in M$, by Proposition I.6-3. We thus see that

$$\langle \nabla'_{X_p} Y, Z_p \rangle = \langle \nabla_{X_p} Y, Z_p \rangle \quad \text{for all } Z_p \in M_p.$$

This is equivalent to the desired result. ♦

2. COROLLARY. If c is a curve in M and Y is a vector field along c which is tangent to M along c , then

$$\frac{DY}{dt} = \mathsf{T} \left(\frac{D'Y}{dt} \right).$$

Consequently, Y is parallel along c , in the sense of parallel that pertains to M , if and only if $D'Y/dt$ is always perpendicular to M . In particular, if Y is parallel along c in the sense that pertains to N , then it is also parallel along c in the sense that pertains to M .

PROOF. There is a unique operation $V \mapsto DV/dt$, from C^∞ vector fields V in M along c to C^∞ vector fields in M along c , with the properties in Proposition II.6-2. From the Theorem it is clear that $V \mapsto \mathsf{T}(DV/dt)$ has these properties. ♦

Merely by combining this information with our previous formulation of other concepts in the ∇ setup, we can immediately deduce further results.

3. COROLLARY. A curve c in M is a geodesic if and only if $D'/dt(dc/dt)$ is everywhere perpendicular to M . In particular, if a geodesic c of N lies entirely in M , then c is also a geodesic in M .

PROOF. The curve c is a geodesic if and only if dc/dt is parallel along c . (The second part also follows from the fact that geodesics are precisely the critical points for the energy function.) ♦

4. COROLLARY. Let M be isometrically immersed in \mathbb{R}^m (with its usual Riemannian metric). A curve c in M is a geodesic if and only if $c''(t)_{c(t)}$ is perpendicular to $M_{c(t)}$ for all t . In particular, a straight line in M is always a geodesic.

PROOF. This is a special case of Corollary 3, for if \mathbb{R}^m has its usual metric, then ∇' is just the directional derivative, so $D'/dt(dc/dt) = c''(t)_{c(t)}$. (In Chapter II.3B we obtained the result (for $m = 3$) by a completely different method.) ♦

Remark: Classically, the tangential component $\mathsf{T}(D'/dt(dc/dt)) = \mathsf{T}(c''(t))$ was called the **geodesic curvature vector** of c . We will meet it again in Chapter 4.

Having considered the tangential component $\mathbb{T}(\nabla'_{X_p} Y)$, it is only fair that we next consider the normal component $\mathbb{L}(\nabla'_{X_p} Y)$. Notice that

$$\mathbb{L}(\nabla'_{X_p} fY) = \mathbb{L}(X_p(f) \cdot Y_p + f(p) \cdot \nabla'_{X_p} Y) = f(p) \cdot \mathbb{L}(\nabla'_{X_p} Y).$$

It follows from our general principal (Theorem I.4-2) that there is a well-defined tensor field s , with $s: M_p \times M_p \rightarrow M_p^\perp$ for each $p \in M$, such that

$$s(X_p, Y_p) = \mathbb{L}(\nabla'_{X_p} Y)$$

for any vector field Y extending Y_p .

5. THEOREM. The tensor s is symmetric.

PROOF. Let X and Y be any extensions of $X_p, Y_p \in M_p$ to all of N which are tangent to M at all points of M . Then

$$\begin{aligned} \mathbb{L}(\nabla'_{X_p} Y) - \mathbb{L}(\nabla'_{Y_p} X) &= \mathbb{L}(\nabla'_{X_p} Y - \nabla'_{Y_p} X) \\ &= \mathbb{L}(\nabla'_X Y(p) - \nabla'_Y X(p)) \\ &= \mathbb{L}([X, Y](p)) = 0, \end{aligned}$$

since $[X, Y]$ is also tangent to M at all points of M (Proposition I.6-3). ♦

By combining Theorems 1 and 5 we can now rewrite the decomposition

$$\nabla'_{X_p} Y = \mathbb{L}(\nabla'_{X_p} Y) + \mathbb{T}(\nabla'_{X_p} Y)$$

in the following form:

The Gauss Formulas:

$$\nabla'_{X_p} Y = \nabla_{X_p} Y + s(X_p, Y_p),$$

where $X_p \in M_p$, and Y is a vector field tangent along M .

Although we seem to be dealing with a single formula here, we will obtain a set of formulas when we choose a coordinate system x^1, \dots, x^n on M and let $X_p = \partial/\partial x^i|_p$ and $Y = \partial/\partial x^j$. Consequently, we will adhere to this classical terminology; it should be compared to the (likewise classical) nomenclature of the next result.

6. THEOREM. Let M be isometrically immersed in N , and let R and R' denote the curvature tensors of M and N , respectively. Then for all $X_p, Y_p, Z_p, W_p \in M_p$ we have

Gauss' Equation (Gauss' Theorema Egregium):

$$\langle R'(X_p, Y_p)Z_p, W_p \rangle = \langle R(X_p, Y_p)Z_p, W_p \rangle + \langle s(X_p, Z_p), s(Y_p, W_p) \rangle - \langle s(Y_p, Z_p), s(X_p, W_p) \rangle$$

PROOF. Extend X_p, Y_p, Z_p, W_p to vector fields X, Y, Z, W which are tangent along M . Then the Gauss formulas yield

$$\begin{aligned} (1) \quad \nabla'_X(\nabla'_Y Z) &= \nabla'_X(\nabla_Y Z) + \nabla'_X(s(Y, Z)) \\ &= \nabla_X(\nabla_Y Z) + s(X, \nabla_Y Z) + \nabla'_X(s(Y, Z)), \end{aligned}$$

and similarly

$$(1') \quad \nabla'_Y(\nabla'_X Z) = \nabla_Y(\nabla_X Z) + s(Y, \nabla_X Z) + \nabla'_Y(s(X, Z)),$$

as well as

$$(2) \quad \nabla'_{[X, Y]} Z = \nabla_{[X, Y]} Z + s([X, Y], Z).$$

Substituting (1), (1'), (2) into the formula $R'(X, Y)Z = \nabla'_X \nabla'_Y Z - \nabla'_Y \nabla'_X Z - \nabla'_{[X, Y]} Z$, and noting that W is orthogonal to any term $s(\cdot, \cdot)$, we obtain

$$(3) \quad \langle R'(X, Y)Z, W \rangle = \langle R(X, Y)Z, W \rangle + \langle \nabla'_X(s(Y, Z)) - \nabla'_Y(s(X, Z)), W \rangle.$$

On the other hand, since $\langle s(Y, Z), W \rangle = 0$ we have

$$\begin{aligned} (4) \quad 0 &= X(\langle s(Y, Z), W \rangle) = \langle \nabla'_X s(Y, Z), W \rangle + \langle s(Y, Z), \nabla'_X W \rangle \\ &= \langle \nabla'_X s(Y, Z), W \rangle + \langle s(Y, Z), \nabla_X W + s(X, W) \rangle \\ &= \langle \nabla'_X s(Y, Z), W \rangle + \langle s(Y, Z), s(X, W) \rangle, \end{aligned}$$

since $\nabla_X W$ is orthogonal to $s(Y, Z)$. The desired result is now obtained by substituting (4), and the similar result with X and Y interchanged, into (3). ♦

Recall that if $P \subset M_p$ is a 2-dimensional subspace of M_p , we define the **sectional curvature** $K(P)$ as $\langle R(X, Y)Y, X \rangle$ for orthonormal $X, Y \in P$. We will let $K'(P)$ denote the corresponding sectional curvature in N .

7. COROLLARY (SYNGE'S INEQUALITY). Let M be isometrically immersed in N , and let $\gamma: [a, b] \rightarrow M$ be a curve in M which is a geodesic in N (and hence also a geodesic in M , by Corollary 3). Then for all 2-dimensional $P \subset M_{\gamma(t)}$ with $\gamma'(t) \in P$ we have

$$K(P) \leq K'(P).$$

In particular, if M is a *surface*, then for all $p = \gamma(t)$ we have

$$K(M_p) \leq K'(M_p).$$

Moreover, in this case equality holds for all $p = \gamma(t)$ if and only if $M_{\gamma(t)}$ is parallel along γ , in the sense that pertains to N .

PROOF. Assume γ is parameterized by arclength. Let $X_p = \gamma'(t)$ and let $Y_p \in P$ be a unit vector perpendicular to X_p . Applying Gauss' equation with $Z_p = Y_p$ and $W_p = X_p$, we obtain

$$K'(P) = K(P) + \langle s(X_p, Y_p), s(X_p, Y_p) \rangle - \langle s(Y_p, Y_p), s(X_p, X_p) \rangle.$$

If we let X be the vector field $X(t) = \gamma'(t)$ along γ , then X is parallel along γ , so we have $0 = \nabla'_X X$. This implies that

$$0 = \perp(\nabla'_X X)(p) = s(X_p, X_p),$$

which gives the desired inequality.

In the case of a surface, we choose $X(t) = \gamma'(t)$ once again, and we let $Y(t)$ be a unit vector in $M_{\gamma(t)}$ which is perpendicular to $X(t)$. Now Gauss' equation gives

$$K'(M_p) = K(M_p) + \langle s(X_p, Y_p), s(X_p, Y_p) \rangle - \langle s(Y_p, Y_p), s(X_p, X_p) \rangle.$$

Once again we have $s(X_p, X_p) = 0$, so equality holds for all p if and only if $s(X_p, Y_p) = 0$ for all p . Moreover, on the *surface* M , the vector field X is parallel along γ , while Y is a unit vector field which makes a constant angle with X along γ . It follows that Y is also parallel along γ , *in the sense that pertains to M* . Therefore

$$0 = \nabla_X Y = \mathbf{T}(\nabla'_X Y).$$

Since

$$s(X_p, Y_p) = \perp(\nabla'_X Y(p)),$$

this shows that $s(X_p, Y_p) = 0$ for all p if and only if $\nabla'_X Y(p) = 0$ for all p ; the latter condition means that $M_{\gamma(t)}$ is parallel along γ . ♦

Remark: We can state the slightly more precise result for M a surface: $K(M_p) = K'(M_p)$ at a particular point p if and only if $\nabla'_X Y(p) = 0$.

As another application of Theorem 6, we give a new proof of an old result: If $W \subset N_p$ is 2-dimensional, and $\mathcal{O} \subset W$ is a sufficiently small neighborhood of 0, then $K(W)$ is the Gaussian curvature at p of the surface $M = \exp(\mathcal{O})$. Clearly we just have to show that $s(X_p, Y_p) = 0$ for $X_p, Y_p \in M_p$, so we just need to show that $s(X_p, X_p) = 0$ for all $X_p \in M_p$. But there is a vector field X tangent to M with $X = c'$ along the geodesic $c(t) = \exp(tX_p)$ of N . Then $\nabla'_X X(p) = 0$, so $s(X_p, X_p) = 0$.

The proof of Corollary 7 has probably already explained why Theorem 6 is referred to in the singular, as “Gauss’ equation”: when M is 2-dimensional and x^1, x^2 is a coordinate system on M , essentially the only interesting case of Theorem 6 occurs for $X_p = W_p = \partial/\partial x^1|_p$ and $Y_p = Z_p = \partial/\partial x^2|_p$, so that we really are dealing with a single equation. This equation actually occurs in Gauss’ paper, as we shall soon see, when we specialize our results somewhat.

For the remainder of this chapter we consider the more specific situation where M is a **hypersurface** in N , that is, a submanifold of codimension 1; we will return only much later to the more general situation. In the case of hypersurfaces we can locally choose a **unit normal field** for M : on a neighborhood U of a point $p \in M$ we can choose a vector field ν such that $\langle \nu, \nu \rangle = 1$ and $\nu(q) \in M_q^\perp$ for all $q \in U$; in fact, there are only two possible choices for ν . Since ν is a vector field of N , defined along M , the symbol $\nabla'_{X_p} \nu$ makes sense for $X_p \in M_p$.

8. THEOREM. Let M be a hypersurface in N , and let ν be a unit normal field on a neighborhood of p in M .

(a) For all $X_p \in M_p$ we have

$$\nabla'_{X_p} \nu \in M_p.$$

(b) If Y is a vector field tangent along M , then we have

The Weingarten Equations:

$$\langle \nabla'_{X_p} \nu, Y_p \rangle = -\langle \nu, \nabla'_{X_p} Y \rangle = -\langle \nu, s(X_p, Y_p) \rangle.$$

(c) Consequently,

$$\langle \nabla'_{X_p} \nu, Y_p \rangle = \langle X_p, \nabla'_{Y_p} \nu \rangle.$$

PROOF. (a) Since $\langle \nu, \nu \rangle = 1$ along M , we have

$$0 = X_p(\langle \nu, \nu \rangle) = 2\langle \nabla'_{X_p} \nu, \nu \rangle,$$

which means that $\nabla'_{X_p} \nu \in M_p$, since M_p^\perp is 1-dimensional.

(b) Since $\langle \nu, Y \rangle = 0$ along M , we have

$$0 = X_p(\langle \nu, Y \rangle) = \langle \nabla'_{X_p} \nu, Y \rangle + \langle \nu, \nabla'_{X_p} Y \rangle,$$

which implies the first equality in the Weingarten equations. The second equality comes from the definition $s(X_p, Y_p) = \perp(\nabla'_{X_p} Y)$, and the fact that $\perp(\nabla'_{X_p} Y)$ is a multiple of ν .

(c) follows from (b) and symmetry of s . ♦

The reader may recall that the “Weingarten equations” have already appeared in Volume II, pg. 124. The relationship between those equations and the ones in Theorem 8, as well as the reason for choosing the notation $s(X_p, Y_p)$, may come out in the following special case of Theorem 8.

9. COROLLARY. Let M^n be a hypersurface in \mathbb{R}^{n+1} and let ν be a unit normal field on a neighborhood of p in M . Then for all $X_p, Y_p \in M_p$ we have

$$s(X_p, Y_p) = \text{II}(X_p, Y_p) \cdot \nu(p),$$

where $\text{II}(X_p, Y_p)$ is the second fundamental form of M defined for the choice ν of unit normal field, namely

$$\text{II}(X_p, Y_p) = -\langle d\nu(X_p), Y_p \rangle.$$

(Here $d\nu(X_p)$ is interpreted as follows [pg. II.121ff.]: Since we can think of ν as a map $\nu: M \rightarrow S^{n-1} \subset \mathbb{R}^{n+1}$, we have the vector-valued differential form $d\nu: M_p \rightarrow \mathbb{R}^{n+1}$, and $d\nu(X_p) \in \mathbb{R}^{n+1}$ is to be moved back to a parallel vector in M_p ; equivalently, $d\nu(X_p)$ denotes $\nu_*(X_p) \in S^{n-1}_{\nu(p)}$ moved back to a parallel vector in M_p .)

PROOF. Since $\nabla'_{X_p} \nu$ is now simply the directional derivative of ν , we have

$$\nabla'_{X_p} \nu = [X_p(\nu)]_p = [d\nu(X_p)]_p = d\nu(X_p),$$

in the notation we have just adopted. So the Theorem says that

$$\begin{aligned} \langle \nu, s(X_p, Y_p) \rangle &= -\langle d\nu(X_p), Y_p \rangle \\ &= \text{II}(X_p, Y_p), \end{aligned}$$

which is equivalent to the desired result. ♦

The reader should now be able to see that the Weingarten equations of Theorem 8 reduce to equations (a)–(c) on pg. II.124 for a surface in \mathbb{R}^3 . More precisely, the equation $\langle \nabla'_{X_p} v, Y_p \rangle = -\langle v, s(X_p, Y_p) \rangle$ establishes the relationship between s and II , and the equation $\langle v, \nabla'_{X_p} Y \rangle = \langle v, s(X_p, Y_p) \rangle$ then reduces to equations (a)–(c). One further point is worth checking: our present proof that s is symmetric is essentially equivalent to our second proof, in Volume II, that II is symmetric.

10. COROLLARY. Let M be a surface immersed in \mathbb{R}^3 , and let $X_p, Y_p \in M_p$. Then

$$\langle R(X_p, Y_p)Y_p, X_p \rangle = \text{II}(X_p, X_p) \cdot \text{II}(Y_p, Y_p) - [\text{II}(X_p, Y_p)]^2.$$

In particular, if (x, y) is a coordinate system on M , and we introduce the classical notation

$$\begin{aligned} \langle \cdot, \cdot \rangle &= \text{I} = E dx \otimes dx + F dx \otimes dy + F dy \otimes dx + G dy \otimes dy \\ \text{II} &= l dx \otimes dx + m dx \otimes dy + m dy \otimes dx + n dy \otimes dy, \end{aligned}$$

then

$$\frac{R_{1212}}{EG - F^2}(p) = \frac{ln - m^2}{EG - F^2}(p) = K(p),$$

where $K(p)$ is the Gaussian curvature of M at p .

PROOF. The first equation follows from Theorem 8, Corollary 9, and the fact that $R' = 0$ for \mathbb{R}^3 . For the second equation we recall the formulas on pp. II.190 and 129. ♦

As we have already noted in the proof of Proposition II.4-7, when we expand R_{1212} using formula (***) on pg. II.188, the second equation in Corollary 10 is exactly equivalent to Gauss' equation for K . The reader probably suspects that our more general Gauss equations can be used to obtain generalizations of the Theorema Egregium to higher dimensions. However, we will defer all such considerations until after we have studied surfaces in more detail. For the present we wish to consider only one more result, which depends on a definition motivated by Corollary 9. If $M \subset N$ is a hypersurface, we produce a symmetric tensor II on M by *defining*

$$s(X_p, Y_p) = \text{II}(X_p, Y_p) \cdot v(p);$$

naturally, the sign of II depends on the choice of the local unit normal field v . Some authors call s the second fundamental form of $M \subset N$, while others

reserve that name for \mathbb{I} . The tensor \mathbb{I} merely gives the length of s up to sign; but since it is real-valued, rather than M_p^\perp valued, the symbol $\nabla_{Z_p}\mathbb{I}$ makes sense. Indeed, we have defined $\nabla_{Z_p}\mathcal{T}$ for any tensor field \mathcal{T} (see Volume II, pp. 229ff.).

11. THEOREM. Let M be a hypersurface in N , and let ν be a unit normal field on a neighborhood of p in M , with corresponding \mathbb{I} . Then for all $X_p, Y_p, Z_p \in M_p$, we have

The Codazzi-Mainardi Equations:

$$\langle R'(X_p, Y_p)Z_p, \nu(p) \rangle = (\nabla_{X_p}\mathbb{I})(Y_p, Z_p) - (\nabla_{Y_p}\mathbb{I})(X_p, Z_p).$$

Remark: This formula gives us the normal component of $R'(X_p, Y_p)Z_p$, while Gauss' equation essentially gives us the tangential component.

PROOF. We begin with the equations derived in the proof of Theorem 6:

$$\begin{aligned} (1) \quad & \nabla'_X(\nabla'_Y Z) = \nabla_X(\nabla_Y Z) + s(X, \nabla_Y Z) + \nabla'_X(s(Y, Z)) \\ (1') \quad & \nabla'_Y(\nabla'_X Z) = \nabla_Y(\nabla_X Z) + s(Y, \nabla_X Z) + \nabla'_Y(s(X, Z)) \\ (2) \quad & \nabla'_{[X, Y]}Z = \nabla_{[X, Y]}Z + s([X, Y], Z) \\ & = \nabla_{[X, Y]}Z + s(\nabla_X Y, Z) - s(\nabla_Y X, Z). \end{aligned}$$

From these we see that the normal component of $R'(X, Y)Z$ is given by

$$\begin{aligned} (3) \quad & \text{normal component of } R'(X, Y)Z = \\ & [\perp \nabla'_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z)] \\ & - [\perp \nabla'_Y(s(X, Z)) - s(\nabla_Y X, Z) - s(X, \nabla_Y Z)]. \end{aligned}$$

On the other hand, since

$$(4) \quad s(Y, Z) = \mathbb{I}(Y, Z) \cdot \nu,$$

we have

$$\nabla'_X(s(Y, Z)) = X(\mathbb{I}(Y, Z)) \cdot \nu + \mathbb{I}(Y, Z) \cdot \nabla'_X \nu,$$

and consequently

$$(5) \quad \langle \nabla'_X(s(Y, Z)), \nu \rangle = X(\mathbb{I}(Y, Z)),$$

since $\nabla'_X v$ is tangent to M . Using (3), (5), and the definition (4) again, we obtain

$$\begin{aligned} \langle R'(X, Y)Z, v \rangle &= [X(\Pi(Y, Z)) - \Pi(\nabla_X Y, Z) - \Pi(Y, \nabla_X Z)] \\ &\quad - [Y(\Pi(X, Z)) - \Pi(\nabla_Y X, Z) - \Pi(X, \nabla_Y Z)]. \end{aligned}$$

The result now follows from* Corollary II.6-5. ❖

It will be useful to examine the form which our fundamental equations take when the ambient space N has constant curvature K_0 . Then by Lemma II.7-18 the curvature tensor R' of N satisfies

$$\begin{aligned} (1) \quad \langle R'(X, Y)Z, W \rangle &= K_0[\langle X, W \rangle \cdot \langle Y, Z \rangle - \langle X, Z \rangle \cdot \langle Y, W \rangle] \\ &\quad \Downarrow \\ (2) \quad R'(X, Y)Z &= K_0[\langle Y, Z \rangle X - \langle X, Z \rangle Y]. \end{aligned}$$

12. COROLLARY. Let N have constant curvature K_0 . Then for M isometrically immersed in N we have

Gauss' Equation:

$$\begin{aligned} \langle R(X_p, Y_p)Z_p, W_p \rangle + \langle s(X_p, Z_p), s(Y_p, W_p) \rangle &- \langle s(Y_p, Z_p), s(X_p, W_p) \rangle \\ &= K_0[\langle X_p, W_p \rangle \cdot \langle Y_p, Z_p \rangle - \langle X_p, Z_p \rangle \cdot \langle Y_p, W_p \rangle]. \end{aligned}$$

And if M is a hypersurface we have

The Codazzi-Mainardi Equations:

$$(\nabla_{X_p} \Pi)(Y_p, Z_p) = (\nabla_{Y_p} \Pi)(X_p, Z_p).$$

PROOF. The first result follows from Theorem 6 and equation (1). The second follows from Theorem 11 and equation (2), which shows that $R'(X_p, Y_p)Z_p$ is tangent to M . ❖

*This Corollary holds also for tensors of type $\binom{k}{0}$; see also Problem 1.

We have carried our analysis of submanifolds as far as we presently wish to go. However, it is also important that we indicate how things work out when we use moving frames. Before doing this, we will first examine the classical tensor analysis treatment of submanifolds. This is included mainly for the sake of completeness, and because you may be unfortunate enough to encounter it again in a classical work which you need to consult. If you are inclined to skip this part, I cannot in good conscience caution you against such a course of action, except to say that reading it must be good for you, because you certainly won't like it.

We will simplify things slightly by beginning with hypersurfaces from the outset. We consider a coordinate system y^1, \dots, y^{n+1} on N , with

$$\langle \ , \ \rangle = \sum_{\alpha, \beta=1}^{n+1} g'_{\alpha\beta} dy^\alpha \otimes dy^\beta,$$

and let x^1, \dots, x^n be a coordinate system on a hypersurface M , so that

$$\langle \ , \ \rangle = \sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j \quad \text{on } M,$$

for certain functions g_{ij} . We adopt the convention that the indices i, j , etc., range from 1 to n , while α, β , etc., range from 1 to $n+1$, even in summation signs, so that \sum_i , for example, denotes $\sum_{i=1}^n$. It is easy to see that

$$\sum_{\alpha, \beta} g'_{\alpha\beta} \frac{\partial y^\alpha}{\partial x^i} \frac{\partial y^\beta}{\partial x^j} = g_{ij} \quad \text{on } M.$$

It will be convenient to let y^α also denote the restriction of y^α to M . Then we can use the symbol $y^\alpha_{;i} = \partial y^\alpha / \partial x^i$, introduced on pg. II.211, for the components of dy^α on M , and we can write

$$(1) \quad \sum_{\alpha, \beta} g'_{\alpha\beta} y^\alpha_{;i} y^\beta_{;j} = g_{ij} \quad \text{on } M.$$

If $v = \sum_{\alpha} v^\alpha \cdot \partial / \partial y^\alpha$ is a unit normal field, then we also have

$$(2) \quad \sum_{\alpha, \beta} g'_{\alpha\beta} y^\alpha_{;i} v^\beta = 0,$$

$$(3) \quad \sum_{\alpha, \beta} g'_{\alpha\beta} v^\alpha v^\beta = 1.$$

We now wish to take the covariant derivative of (1) on M . Notice that $y^\alpha{}_{;i} y^\beta{}_{;j}$ is the (i, j) component of the tensor $dy^\alpha \otimes dy^\beta$ on M ; on the other hand, each $g'_{\alpha\beta}$ is just a function on M . Writing

$$\frac{\partial g'_{\alpha\beta}}{\partial x^k} \quad \text{as} \quad \sum_{\gamma} \frac{\partial g'_{\alpha\beta}}{\partial y^\gamma} \frac{\partial y^\gamma}{\partial x^k} = \sum_{\gamma} \frac{\partial g'_{\alpha\beta}}{\partial y^\gamma} y^\gamma{}_{;k},$$

and using Proposition II.5-2, we obtain from (1)

$$\begin{aligned} \sum_{\alpha, \beta, \gamma} \frac{\partial g'_{\alpha\beta}}{\partial y^\gamma} y^\alpha{}_{;i} y^\beta{}_{;j} y^\gamma{}_{;k} + \sum_{\alpha, \beta} g'_{\alpha\beta} (y^\alpha{}_{;ik} y^\beta{}_{;j} + y^\beta{}_{;jk} y^\alpha{}_{;i}) &= g_{ij;k} \\ &= 0, \quad \text{by Ricci's Lemma (Proposition II.5-3).} \end{aligned}$$

If we write this equation with i and k interchanged, the term

$$\sum_{\alpha, \beta, \gamma} \frac{\partial g'_{\alpha\beta}}{\partial y^\gamma} y^\alpha{}_{;k} y^\beta{}_{;j} y^\gamma{}_{;i} \quad \text{can be replaced by} \quad \sum_{\alpha, \beta, \gamma} \frac{\partial g'_{\gamma\beta}}{\partial y^\alpha} y^\alpha{}_{;i} y^\beta{}_{;j} y^\gamma{}_{;k}.$$

A similar replacement can be made when we rewrite the original equation with j and k interchanged. Adding the two equations so obtained, and subtracting the original, we get

$$\sum_{\alpha, \beta} g'_{\alpha\beta} y^\alpha{}_{;k} y^\beta{}_{;ij} + \sum_{\alpha, \beta, \gamma} [\alpha\beta, \gamma]' y^\alpha{}_{;i} y^\beta{}_{;j} y^\gamma{}_{;k} = 0,$$

where $[\ , \]'$ indicates the Christoffel symbols for the y coordinate system. This can also be written as

$$(4) \quad \sum_{\alpha, \beta} g'_{\alpha\beta} y^\beta{}_{;k} \left(y^\alpha{}_{;ij} + \sum_{\rho, \sigma} \Gamma'^{\alpha}_{\rho\sigma} y^\rho{}_{;i} y^\sigma{}_{;j} \right) = 0,$$

which shows that the expression in parentheses is the α component of a vector perpendicular to M . As a matter of fact, a calculation (Problem 2) shows that it is the coefficient of $\partial/\partial y^\alpha$ in the expression for $\nabla'_{\partial/\partial x^i} \partial/\partial x^i - \nabla_{\partial/\partial x^i} \partial/\partial x^i$. Consequently, (4) is equivalent to Theorem 1, and despite the ugliness of the equations involved, its derivation is clearly closely related to that of Theorem 1.

Since $\sum_{\alpha} v^\alpha \cdot \partial/\partial y^\alpha$ is a unit normal field, and M_p^\perp has dimension 1, equation (4) implies that

$$(5) \quad y^\alpha{}_{;ij} + \sum_{\rho, \sigma} \Gamma'^{\alpha}_{\rho\sigma} y^\rho{}_{;i} y^\sigma{}_{;j} = \Pi_{ij} v^\alpha$$

for certain functions Π_{ij} ; multiplying by $\sum_{\beta} g'_{\alpha\beta} v^{\beta}$ and using (3), this can be written

$$(6) \quad \Pi_{ij} = \sum_{\alpha, \beta} g'_{\alpha\beta} v^{\beta} y^{\alpha}_{;ij} + \sum_{\rho, \sigma, \beta} [\rho\sigma, \beta]' y^{\rho}_{;i} y^{\sigma}_{;j} v^{\beta}.$$

This shows that the Π_{ij} satisfy the transformation rule for a covariant tensor of order 2 on M , since the $y^{\alpha}_{;ij}$ and $y^{\rho}_{;i} y^{\sigma}_{;j}$ do, and since the other terms don't involve the coordinate system x but only the coordinate system y on N (at the same time we see that the whole right side doesn't even depend on y). It is also clear that $\Pi_{ij} = \Pi_{ji}$. Equation (5) is equivalent to the Gauss formulas.

We next take the covariant derivative of equation (2) on M (now both $g'_{\alpha\beta}$ and v^{β} are functions on M). We obtain

$$\begin{aligned} \sum_{\alpha, \beta} g'_{\alpha\beta} (y^{\alpha}_{;ij} v^{\beta} + y^{\alpha}_{;i} v^{\beta}_{;j}) &= - \sum_{\alpha, \beta, \sigma} y^{\alpha}_{;i} y^{\sigma}_{;j} v^{\beta} \frac{\partial g'_{\alpha\beta}}{\partial y^{\sigma}} \\ &= - \sum_{\alpha, \beta, \sigma} y^{\alpha}_{;i} y^{\sigma}_{;j} v^{\beta} ([\alpha\sigma, \beta]' + [\beta\sigma, \alpha]'). \end{aligned}$$

Then (6) gives

$$\Pi_{ij} = - \sum_{\alpha, \beta} g'_{\alpha\beta} y^{\alpha}_{;i} v^{\beta}_{;j} - \sum_{\rho, \sigma, \beta} [\beta\sigma, \rho]' y^{\rho}_{;i} y^{\sigma}_{;j} v^{\beta},$$

or

$$-\Pi_{ij} = \sum_{\alpha, \beta} g'_{\alpha\beta} y^{\alpha}_{;i} \left(v^{\beta}_{;j} + \sum_{\rho, \sigma} \Gamma'^{\beta}_{\rho\sigma} y^{\rho}_{;j} v^{\sigma} \right),$$

which can also be written as

$$(7) \quad -\Pi_{ij} = \sum_{\alpha, \beta} g'_{\alpha\beta} y^{\alpha}_{;i} \left(\sum_{\rho} v^{\beta}_{;\rho} y^{\rho}_{;j} \right),$$

where $v^{\beta}_{;\rho}$ now denotes the covariant derivative, in N , of the vector field with components v^{β} , so that

$$v^{\beta}_{;\rho} = \frac{\partial v^{\beta}}{\partial y^{\rho}} + \sum_{\sigma} \Gamma'^{\beta}_{\rho\sigma} v^{\sigma}.$$

Equation (7) is clearly equivalent to one part of the Weingarten equations, namely $-\langle v, s(X_p, Y_p) \rangle = \langle \nabla'_{X_p} v, Y_p \rangle$. If we treat (3) in a similar manner, we end up with

$$(8) \quad \sum_{\alpha, \beta} g'_{\alpha\beta} v^{\alpha} \left(\sum_{\rho} v^{\beta}_{;\rho} y^{\rho}_{;j} \right) = 0,$$

which is equivalent to the fact that $\nabla'_{X_p} v \in M_p$.

Since (8) shows that we can write

$$(9) \quad \sum_{\rho} v^{\beta}{}_{;\rho} y^{\rho}{}_{;j} = \sum_k A_j^k y^{\beta}{}_{;k} \quad \text{for some functions } A_j^k \text{ on } M,$$

equation (7) gives

$$\begin{aligned} -\Pi_{ij} &= \sum_{\alpha, \beta, k} g'_{\alpha\beta} y^{\alpha}{}_{;i} y^{\beta}{}_{;k} A_j^k \\ &= \sum_k g_{ik} A_j^k \quad \text{by (I),} \end{aligned}$$

so

$$A_j^m = \sum_l -\Pi_{lj} g^{lm},$$

and hence from (9)

$$\sum_{\rho} v^{\beta}{}_{;\rho} y^{\rho}{}_{;j} = - \sum_{l,m} \Pi_{lj} g^{lm} y^{\beta}{}_{;m},$$

or equivalently

$$(10) \quad v^{\beta}{}_{;j} + \sum_{\rho, \sigma} \Gamma'_{\rho\sigma}{}^{\beta} y^{\rho}{}_{;j} v^{\sigma} = - \sum_{l,m} \Pi_{lj} g^{lm} y^{\beta}{}_{;m}.$$

We return to equation (5), equivalent to Gauss' formulas. We can apply Ricci's identity (Proposition II.5-4)

$$\lambda_{i;jk} - \lambda_{i;kj} = \sum_m \lambda_m R^m{}_{ijk} = \sum_{m,h} \lambda_m g^{mh} R_{hijk}$$

to $\lambda_i = y^{\alpha}{}_{;i}$, obtaining

$$\sum_{m,h} y^{\alpha}{}_{;m} g^{mh} R_{hijk} = y^{\alpha}{}_{;ijk} - y^{\alpha}{}_{;ikj}.$$

Computing the $y^{\alpha}{}_{;ijk}$ from (5), and using (5) and (10) in the result, we obtain finally

$$\begin{aligned} \sum_{m,h} y^{\alpha}{}_{;m} g^{mh} [R_{hijk} - (\Pi_{hj} \Pi_{ik} - \Pi_{hk} \Pi_{ij})] - v^{\alpha} (\Pi_{ij;k} - \Pi_{ik;j}) \\ - \sum_{\rho, \sigma, \lambda} R'^{\alpha}{}_{\rho\sigma\lambda} y^{\rho}{}_{;i} y^{\sigma}{}_{;j} y^{\lambda}{}_{;k} = 0. \end{aligned}$$

Multiplying by $\sum_{\alpha} g'_{\alpha\beta} y^{\beta}_{;l}$ or by $\sum_{\alpha} g'_{\alpha\beta} v^{\beta}$, we get

$$(11) \quad R_{ijkl} = (\Pi_{ik}\Pi_{jl} - \Pi_{il}\Pi_{jk}) + \sum_{\alpha,\beta,\gamma,\delta} R'_{\alpha\beta\gamma\delta} y^{\alpha}_{;i} y^{\beta}_{;j} y^{\gamma}_{;k} y^{\delta}_{;l}$$

$$(12) \quad \Pi_{ij;k} - \Pi_{ik;j} = \sum_{\alpha,\beta,\gamma,\delta} R'_{\alpha\beta\gamma\delta} y^{\alpha}_{;i} y^{\gamma}_{;j} y^{\delta}_{;k} v^{\beta}.$$

Equations (11) and (12) are equivalent to Gauss' Equation and the Codazzi-Mainardi equations, respectively. Whew!

When we turn to the method of moving frames, we find ourselves in a situation completely different from the mass of calculations in which we have just been mired. Although the moving frame method will not have the geometric appeal of the ∇ theory, it is far superior computationally. Not only are all the equations short and simple, but all the results follow naturally, almost without thought, from the structural equations. Indeed, everything happens so quickly that the real problem is recognizing a result when it appears.

Consider first an orthonormal moving frame X_1, \dots, X_n on an open subset of M . Recall that the **dual 1-forms** θ^i are defined by $\theta^i(X_j) = \delta^i_j$, and that there are unique 1-forms ω^i_j , the **connection forms**, satisfying the two equations

$$(1) \quad \omega^i_j = -\omega^j_i$$

$$(2) \quad d\theta^i = -\sum_{k=1}^n \omega^i_k \wedge \theta^k \quad (\text{The first structural equation}).$$

The **curvature forms** Ω^i_j are then defined by

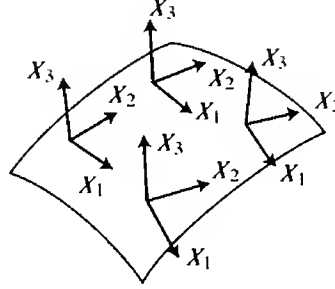
$$(3) \quad d\omega^i_j = -\sum_{k=1}^n \omega^i_k \wedge \omega^k_j + \Omega^i_j \quad (\text{The second structural equation}).$$

The relationship between ω^i_j, Ω^i_j and ∇, R is given by

$$(4) \quad \nabla_{X_k} X_j = \sum_{i=1}^n \omega^i_j(X_k) X_i \quad \text{or} \quad \langle \nabla_X X_j, X_i \rangle = \omega^i_j(X)$$

$$(5) \quad R(X_k, X_l)X_j = \sum_{i=1}^n \Omega^i_j(X_k, X_l) X_i \quad \text{or} \quad \langle R(X, Y)X_j, X_i \rangle = \Omega^i_j(X, Y).$$

Now let us consider an orthonormal moving frame X_1, \dots, X_m , defined on an open subset of N , with the property that X_1, \dots, X_n are tangent to M at points of M , and consequently X_{n+1}, \dots, X_m are normal to M at points of M ; such an orthonormal moving frame is said to be **adapted to M** . An adapted



orthonormal moving frame gives us an orthonormal moving frame X_1, \dots, X_n along M , with corresponding forms $\theta^i, \omega_j^i, \Omega_j^i$ ($i, j \leq n$). We also want to consider the corresponding forms for the entire moving frame X_1, \dots, X_m on N ; these will be denoted by $\phi^\alpha, \psi_\beta^\alpha, \Psi_\beta^\alpha$. We adopt the convention that i, j , etc., always range from 1 to n , while α, β , etc., always range from 1 to m , even in summation signs, so that \sum_i , for example, means $\sum_{i=1}^n$; it will also be convenient to use r, s , etc., for numbers that range from $n+1$ to m .

Now the forms $\phi^\alpha, \psi_\beta^\alpha, \Psi_\beta^\alpha$ can be restricted to TM (that is, to tangent vectors of M). Clearly

$$\begin{aligned} \phi^i &= \theta^i & \text{on } TM & \quad i \leq n \\ \phi^r &= 0 & \text{on } TM & \quad r > n. \end{aligned}$$

To obtain some information about the forms ψ_β^α on TM , we look at the first structural equation,

$$d\phi^\alpha = - \sum_\gamma \psi_\gamma^\alpha \wedge \phi^\gamma.$$

Restricting to TM we obtain

$$\begin{aligned} \text{(a)} \quad d\theta^i &= - \sum_k \psi_k^i \wedge \theta^k & \text{on } TM & \quad i \leq n \\ \text{(b)} \quad 0 &= \sum_k \theta^k \wedge \psi_k^r & \text{on } TM & \quad r > n. \end{aligned}$$

Recall that ω_j^i were the *unique* forms satisfying (1), (2). Since $\psi_\beta^\alpha = -\psi_\alpha^\beta$, equation (a) therefore shows that

$$\text{(c)} \quad \psi_j^i = \omega_j^i \quad \text{on } TM \quad i, j \leq n.$$

This equation already contains some information! In fact, equation (4) shows that

$$\langle \nabla_X X_j, X_i \rangle = \omega_j^i(X) = \psi_j^i(X) = \langle \nabla'_X X_j, X_i \rangle \quad \text{for } X \in TM;$$

this is exactly equivalent to Theorem 1, for it shows that $\nabla'_X X_j(p)$ has the same inner product with every element of M_p as $\nabla_X X_j(p) \in M_p$.

Next, we look at the forms $\psi_k^r = -\psi_r^k$ on TM , for $k \leq n < r$. Equation (b) tells us that they satisfy the hypothesis of the following Lemma.

13. LEMMA (CARTAN'S LEMMA). If $\lambda^1, \dots, \lambda^n$ are (C^∞) linearly independent 1-forms on a manifold M (of dimension $n' \geq n$), and μ_1, \dots, μ_n are (C^∞) 1-forms on M satisfying

$$(*) \quad \sum_{i=1}^n \lambda^i \wedge \mu_i = 0,$$

then there are unique (C^∞) functions f_{ij} on M such that

$$\mu_i = \sum_{j=1}^n f_{ij} \lambda^j;$$

moreover,

$$f_{ij} = f_{ji}.$$

Remark: This result (or at least the corresponding result for vector spaces) has already been given in Problem I.7-11.

PROOF. In a neighborhood of any point we can choose (C^∞) 1-forms $\lambda^{n+1}, \dots, \lambda^{n'}$ so that $\lambda^1, \dots, \lambda^{n'}$ are everywhere independent. Then there are (C^∞) functions f_{ij} ($i \leq n, j \leq n'$) with

$$\mu_i = \sum_{j=1}^{n'} f_{ij} \lambda^j.$$

Equation $(*)$ implies that

$$0 = \sum_{i=1}^n \sum_{j=1}^{n'} f_{ij} \lambda^i \wedge \lambda^j = \sum_{1 \leq i < j \leq n} (f_{ij} - f_{ji}) \lambda^i \wedge \lambda^j + \sum_{i=1}^n \sum_{j > n} f_{ij} \lambda^i \wedge \lambda^j.$$

Since the $\lambda^i \wedge \lambda^j$ for $i < j$ are linearly independent, we have $f_{ij} - f_{ji} = 0$ for $i, j \leq n$ and $f_{ij} = 0$ for $j > n$. ♦

Applying Cartan's Lemma to the ψ_k^r , we conclude that there are unique functions s_{ij}^r on M satisfying

$$(d) \quad \begin{aligned} \psi_j^r &= -\psi_r^j = \sum_i s_{ij}^r \theta^i \quad \text{on } TM \quad r > n, \\ s_{ij}^r &= s_{ji}^r. \end{aligned}$$

Equation (4) now shows that

$$\langle \nabla'_{X_k} X_j, X_r \rangle = \psi_j^r(X_k) = s_{kj}^r \quad r > n,$$

and hence

$$\langle \nabla'_{X_j} X_k, X_r \rangle = \langle \nabla'_{X_k} X_j, X_r \rangle \quad r > n.$$

This is basically Theorem 5, asserting the symmetry of s , which in our present notation can be defined by setting

$$(e) \quad s(X_j, X_k) = \sum_r \psi_j^r(X_k) \cdot X_r = \sum_r s_{jk}^r X_r,$$

and extending s by linearity. It should be noted that this definition of s involves a choice of a moving frame; when one is developing everything from the moving frame approach, a little calculation (Problem 3) must be supplied to show that the definition of s is really independent of the choice. Equations (c) and (d) together are equivalent to the Gauss formulas.

Now let us look at the second structural equation

$$d\psi_\beta^\alpha = - \sum_\gamma \psi_\gamma^\alpha \wedge \psi_\beta^\gamma + \Psi_\beta^\alpha.$$

Restricting to TM we obtain, for $\alpha, \beta = i, j \leq n$,

$$(f) \quad d\omega_j^i = - \sum_k \omega_k^i \wedge \omega_j^k + \sum_r \psi_r^i \wedge \psi_j^r + \Psi_j^i \quad \text{on } TM \quad i, j \leq n.$$

Comparing with (3) we obtain

$$(g) \quad \Psi_j^i = \Omega_j^i - \sum_r \psi_r^i \wedge \psi_j^r \quad \text{on } TM.$$

Then equation (5) gives

$$\begin{aligned} \langle R'(X, Y)X_j, X_i \rangle &= \Psi_j^i(X, Y) = \langle R(X, Y)X_j, X_i \rangle \\ &\quad - \sum_r (\psi_r^i(X) \psi_j^r(Y) - \psi_r^i(Y) \psi_j^r(X)). \end{aligned}$$

Since we have, for example,

$$\begin{aligned}\sum_r \psi_i^r(X) \psi_j^r(Y) &= \sum_r \langle s(X_i, X), X_r \rangle \cdot \langle s(X_j, Y), X_r \rangle \\ &= \langle s(X_i, X), s(X_j, Y) \rangle,\end{aligned}$$

it follows that (g) is exactly equivalent to Theorem 6 (Gauss' Equation).

If we instead choose $\alpha = r > n$, and $\beta = j \leq n$, we obtain

$$(h) \quad d\psi_j^r = - \sum_i \psi_i^r \wedge \omega_j^i - \sum_s \psi_s^r \wedge \psi_j^s + \Psi_j^r \quad \text{on } TM.$$

As before, we now restrict ourselves to the case $m = n + 1$; then X_{n+1} is a unit normal field on M . Notice that the equation $\psi_{n+1}^j = -\psi_j^{n+1}$ gives

$$\langle \nabla'_{X_k} X_{n+1}, X_j \rangle = \psi_{n+1}^j(X_k) = -\psi_j^{n+1}(X_k) = -\langle X_{n+1}, s(X_j, X_k) \rangle,$$

which are the Weingarten equations. Equation (h) takes the form

$$d\psi_j^{n+1} = - \sum_i \psi_i^{n+1} \wedge \omega_j^i + \Psi_j^{n+1}.$$

A little work (Problem 4) shows that this is equivalent to the Codazzi-Mainardi equations.

SUMMARY

$$\begin{array}{l} \phi^i = \theta^i \quad \text{on } TM \\ \phi^r = 0 \quad \text{on } TM \end{array}$$

$$\text{Consequences of the first structural equation} \left\{ \begin{array}{l} \psi_j^i = \omega_j^i \quad \text{on } TM \dots\dots\dots (\text{Theorem 1}) \\ \psi_j^r = \sum_i s_{ij}^r \theta^i \quad \text{on } TM \dots\dots\dots (\text{Theorem 6}) \\ s_{ij}^r = s_{ji}^r \end{array} \right\} \begin{array}{l} \text{The} \\ \text{Gauss} \\ \text{formulas} \end{array}$$

$$\text{Consequences of the second structural equation} \left\{ \begin{array}{l} \Psi_j^i = \Omega_j^i - \sum_r \psi_i^r \wedge \psi_j^r \quad \text{on } TM \dots\dots \text{Gauss' Equation} \\ \text{For } m = n + 1: \\ \Psi_j^{n+1} = d\psi_j^{n+1} + \sum_i \psi_i^{n+1} \wedge \omega_j^i \dots\dots \text{Codazzi-Mainardi} \\ \hspace{10em} \text{on } TM \hspace{1em} \text{Equations} \end{array} \right.$$

Although the derivation of the fundamental equations was so much easier in terms of moving frames than in terms of tensors, the resultant equations have the same disadvantage as the moving frame treatment of connections itself—our equations are not “invariant”, they are merely a set of equations which hold for each choice of adapted orthonormal moving frame. Moreover, it is very hard to get any geometric feel for the equations—the tensor form of the equations seem much more geometric, especially Gauss’ equation. As one might guess, an invariant description of the moving frame equations can be obtained by considering an appropriate principal bundle—the “bundle of adapted orthonormal frames”. In Chapter 7 we will actually consider this construction in detail, even for submanifolds of higher codimension, but we will do this mainly for the sake of completeness, since the results which we will derive from this construction will also be obtained in other ways. On the whole, it seems uneconomical to construct the elaborate machinery of a principal bundle just to have a gadget on which we can give an invariant formulation of the fundamental equations for submanifolds, especially since the invariant equations are even more abstract and ungeometric. It is much easier, and more satisfying, to state these equations in terms of tensors down on the submanifold itself. On the other hand, when it comes to *using* these equations to prove theorems about submanifolds, the equations in terms of moving frames will almost always prove to be superior.

ADDENDUM

AUTO-PARALLEL AND TOTALLY GEODESIC SUBMANIFOLDS

The material of this section, besides being of interest in its own right, will play an important role on several occasions later on. We will not require any tools not already developed within the chapter, even though we will be dealing with submanifolds of arbitrary codimension, and even with manifolds whose connection does not come from a Riemannian metric.

Let N be a manifold with a connection ∇' , and let M be a submanifold of N . We say that M is **auto-parallel** if the parallel translation in N along a curve c in M always takes vectors tangent to M into vectors tangent to M . For example, a straight line or a plane in \mathbb{R}^3 is auto-parallel.

14. PROPOSITION. A submanifold M of (N, ∇') is auto-parallel if and only if $\nabla'_X Y$ is tangent to M whenever X and Y are.

PROOF. We know from Proposition II.6-3 that

$$\nabla'_{X_p} Y = \lim_{h \rightarrow 0} \frac{1}{h} (\tau_h^{-1} Y_{c(h)} - Y_p),$$

where c is a curve with $c'(0) = X_p$, and τ_h is parallel translation along c from $c(0)$ to $c(h)$. This makes it immediately clear that if M is auto-parallel, then $\nabla'_{X_p} Y$ is tangent to M if X and Y are.

Conversely, suppose $\nabla'_{X_p} Y$ is tangent to M whenever X and Y are. Let c be a curve in M and let V be a parallel vector field along c . Choose a coordinate system $x^1, \dots, x^n, x^{n+1}, \dots, x^m$ for N such that $x^r = 0$ on M for all $r > n$. If $V_t \in N_{c(t)}$ is given by

$$V_t = \sum_{\alpha=1}^m v^\alpha(t) \cdot \frac{\partial}{\partial x^\alpha} \Big|_c(t),$$

then we have (pg. II.233)

$$(1) \quad 0 = \frac{dv^\gamma(t)}{dt} + \sum_{\alpha, \beta} \frac{dc^\alpha(t)}{dt} \Gamma_{\alpha\beta}^\gamma(c(t)) v^\beta(t).$$

Now $c^r(t) = 0$ for $r > n$, since c lies in M . Moreover, if $i, j \leq n$, then the vector

$$\nabla' \frac{\partial}{\partial x^i} \frac{\partial}{\partial x^j} = \sum_{\gamma} \Gamma_{ij}^{\gamma} \frac{\partial}{\partial x^{\gamma}}$$

is tangent to M by hypothesis, so we must have $\Gamma_{ij}^r = 0$ for $r > n$. So for $\gamma = s > n$, equation (1) becomes

$$\frac{dv^s(t)}{dt} = - \sum_{r=n+1}^m \sum_{i=1}^n \frac{dc^i(t)}{dt} \Gamma_{ir}^s(c(t)) v^r(t).$$

This set of $m - n$ equations for the $m - n$ functions v^s has a unique solution for a given initial condition. The solution with all $v^s(0) = 0$ is clearly just $v^s(t) = 0$ for all $s > n$. In other words, if V_0 is tangent to M , then so are all V_t . ♦

15. COROLLARY. If M is an auto-parallel submanifold of (N, ∇') , then

$$R'(X_p, Y_p)Z_p \in M_p \quad \text{for all } X_p, Y_p, Z_p \in M_p.$$

PROOF. Use the definition

$$R'(X, Y)Z = \nabla'_X \nabla'_Y Z - \nabla'_Y \nabla'_X Z - \nabla'_{[X, Y]} Z$$

(and Proposition I.6-3). ♦

In the particular case where the connection ∇' on N is the unique symmetric connection compatible with a Riemannian metric $\langle \cdot, \cdot \rangle$ on N , we can characterize auto-parallel submanifolds $M \subset N$ in a different way.

16. PROPOSITION. If $(N, \langle \cdot, \cdot \rangle)$ is a Riemannian manifold, then a submanifold $M \subset N$ is auto-parallel if and only if the second fundamental form s of M is zero.

PROOF. By definition, s is zero if and only if $\nabla'_X Y$ is tangent to M whenever X and Y are. So the result follows from Proposition 14. ♦

Notice that whenever M is an auto-parallel submanifold of (N, ∇') , we can define a connection ∇ on M by letting $\nabla_X Y = \nabla'_X Y$ for X and Y tangent to M . This connection ∇ on M is called the **induced connection** on M . (In the

Riemannian case, ∇ clearly coincides with the connection M has as a Riemannian submanifold.) If c is a curve in M and V is a vector field along c which is everywhere tangent to M , then the covariant derivative DV/dt along c which is determined by ∇ is exactly the same as the covariant derivative $D'V/dt$ along c which is determined by ∇' : for the proof we just apply Proposition II.6-2, which essentially defines DV/dt . In particular, if V is parallel along c with respect to the connection ∇ in M , then it is also parallel along c with respect to the connection ∇' in N .

Auto-parallel submanifolds can also be characterized in yet another way. A submanifold M of (N, ∇') is called **geodesic at p** if every geodesic γ with $\gamma(0) = p$ and $\gamma'(0) \in M_p$ remains in M on some interval $(-\varepsilon, \varepsilon)$. It is called **totally geodesic** if it is geodesic at every point; it is easy to see that $M \subset N$ is totally geodesic if and only if every geodesic in M is also a geodesic in N .

17. PROPOSITION. Let M be a submanifold of a manifold (N, ∇') .

- (1) If M is auto-parallel, then M is totally geodesic.
- (2) If M is totally geodesic, and ∇' is symmetric, then M is auto-parallel.

PROOF. (1) Let c be a geodesic of N with $c'(0) \in M_p$. Let \bar{c} be the geodesic in M , with respect to the induced connection ∇ , satisfying $\bar{c}'(0) = c'(0)$. To prove that an interval of c lies in M , it certainly suffices to prove that $\bar{c} = c$ along some interval containing 0. Now by the definition of a geodesic, $d\bar{c}/dt$ is parallel along c with respect to ∇ . As we have already noted, this implies that $d\bar{c}/dt$ is parallel along c with respect to ∇' . Thus \bar{c} is a geodesic in N . Since $\bar{c}'(0) = c'(0)$, the geodesics \bar{c} and c must coincide on their common domain.

(2) In a neighborhood of a point $p \in M$ we can choose a coordinate system $x^1, \dots, x^n, x^{n+1}, \dots, x^m$ for N such that $x^r = 0$ on M for $r > n$. Let c be a geodesic with

$$c(0) = p \quad \text{and} \quad c'(0) = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \Big|_p \in M_p.$$

Then by hypothesis we have $c(t) \in M$ for sufficiently small t . Now c satisfies (pg. II.246)

$$\frac{d^2 c^\gamma}{dt^2} + \sum_{\alpha, \beta} \Gamma_{\alpha\beta}^\gamma(c(t)) \frac{dc^\alpha}{dt} \frac{dc^\beta}{dt} = 0.$$

For $\gamma = s > n$ we have

$$\sum_{i,j=1}^n \Gamma_{ij}^s(c(t)) \frac{dc^i}{dt} \frac{dc^j}{dt} = 0 \quad \text{for small } t.$$

Letting $t = 0$, we obtain

$$\sum_{i,j=1}^n \Gamma_{ij}^s(p) a^i a^j = 0.$$

Choosing $a^i = 1$, all other $a^j = 0$, we get

$$(a) \quad 0 = \Gamma_{ii}^s(p).$$

Choosing $a^i = a^j = 1$, all other $a^k = 0$, we get

$$(b) \quad 0 = \Gamma_{ii}^s(p) + \Gamma_{ij}^s(p) + \Gamma_{ji}^s(p) + \Gamma_{jj}^s(p) = \Gamma_{ij}^s(p) + \Gamma_{ji}^s(p).$$

Using symmetry of the Γ 's, we find that $\Gamma_{ij}^s(p) = 0$ for all i, j . This is true for all $p \in M$, so we find that $\nabla'_X Y$ is tangent to M if X and Y are. The result then follows from Proposition 14. ♦

Every n -dimensional plane $P \subset \mathbb{R}^m$ is clearly totally geodesic. (Conversely, if $M \subset \mathbb{R}^m$ is a totally geodesic submanifold, and p is a point of M , then M must clearly contain a neighborhood of its tangent space $M_p \subset \mathbb{R}^m$; so if M is a connected n -dimensional totally geodesic submanifold of \mathbb{R}^m , then M must be part of an n -dimensional plane $P \subset \mathbb{R}^m$.) It is just as clear that if we give S^m its standard Riemannian metric, then any n -sphere $S^n \subset S^m$ is totally geodesic. Now let us consider the Riemannian manifold $(N, \langle \ , \ \rangle)$ mentioned on pg. II.301, with constant curvature -1 : the manifold N is

$$N = \left\{ a \in \mathbb{R}^m : \sum_{\alpha=1}^m (a^\alpha)^2 < 4 \right\},$$

and the components $g_{\alpha\beta}$ of $\langle \ , \ \rangle$ with respect to the usual coordinate system x^1, \dots, x^m are given by

$$g_{\alpha\beta} = \frac{\delta_{\alpha\beta}}{\left[1 - \frac{1}{4} \sum_{\alpha=1}^m (x^\alpha)^2 \right]^2}.$$

Let $M \subset N$ be

$$M = \{a \in N : a^{n+1} = \dots = a^m = 0\}.$$

The formulas for the Γ 's on pg. II.299 show that $\Gamma_{ij}^r = 0$ on M whenever $i, j \leq n$ and $r > n$, which means that $\nabla'_X Y$ is tangent to M whenever X and Y are. So M is a totally geodesic submanifold of N , by Propositions 14 and 16. Since the metric $\langle \cdot, \cdot \rangle$ is radially symmetric around 0, it is clear that we can find a totally geodesic submanifold M of N with M_0 being any n -dimensional subspace of N_0 . The same is true at any other point $p \in N$, because the fact that N has constant curvature implies that p has a neighborhood isometric to a neighborhood of 0 (Corollary II.7-13); in fact, since N is simply-connected and complete, there is actually an isometry of N onto itself taking any point p to 0 (Problem 5).*

The possibility of finding so many totally geodesic submanifolds is very exceptional:

18. THEOREM. Let $(N, \langle \cdot, \cdot \rangle)$ be a connected Riemannian manifold of dimension $m \geq 3$. Suppose that for all $p \in N$ and all 2-dimensional subspaces $P \subset N_p$ there is a totally geodesic submanifold M of N with $p \in M$ and $M_p = P$. Then N has constant curvature.

PROOF. Each submanifold M is auto-parallel, by Proposition 17, so Corollary 15 shows that

$$\langle R'(X_p, Y_p)Z_p, W_p \rangle = 0$$

for $X_p, Y_p, Z_p \in M_p$ and $W_p \in M_p^\perp$. Since M_p can be any 2-dimensional subspace $P \subset N_p$, we see that

$$(1) \quad \langle R'(X_p, Y_p)X_p, W_p \rangle = 0 \quad \text{for orthonormal } X_p, Y_p, W_p \in N_p.$$

Applying (1) to $X_p, \bar{Y}_p, \bar{W}_p$ with

$$\begin{aligned} \bar{Y}_p &= (\cos \alpha)Y_p + (\sin \alpha)W_p \\ \bar{W}_p &= (-\sin \alpha)Y_p + (\cos \alpha)W_p, \end{aligned}$$

we obtain

$$\begin{aligned} 0 &= \sin \alpha \cos \alpha [\langle R'(X_p, W_p)X_p, W_p \rangle - \langle R'(X_p, Y_p)X_p, Y_p \rangle] \\ &\quad + \cos^2 \alpha \langle R'(X_p, Y_p)X_p, W_p \rangle - \sin^2 \alpha \langle R'(X_p, W_p)X_p, Y_p \rangle \\ &= \sin \alpha \cos \alpha [\langle R'(X_p, W_p)X_p, W_p \rangle - \langle R'(X_p, Y_p)X_p, Y_p \rangle] \quad \text{by (1).} \end{aligned}$$

*More detailed information about the manifold N will be found in Chapter 7, Part A.

Thus $\langle R'(X_p, W_p)X_p, W_p \rangle = \langle R'(X_p, Y_p)X_p, Y_p \rangle$ for all orthonormal X_p, Y_p, W_p , which implies that all sectional curvatures at p are equal. Since this is true for all p , Schur's Theorem (II.7-19) shows that M has constant curvature. ❖

It seems rather clear that if one takes a Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$ “at random”, then it will not have any totally geodesic submanifolds of dimension > 1 . But I must admit that I don't know of any specific example of such a manifold.

PROBLEMS

1. (a) In Corollary II.6-5, each $A(p)$ is regarded as a map $M_p \times \cdots \times M_p \rightarrow M_p$. If we instead regard each $A(p)$ as a map $M_p \times \cdots \times M_p \times M_p^* \rightarrow \mathbb{R}$, show that

$$\begin{aligned} (\nabla_{X_p} A)(Y_1(p), \dots, Y_k(p), \omega(p)) &= \nabla_{X_p}(A(Y_1, \dots, Y_k, \omega)) \\ &- \sum_{i=1}^k A(Y_1(p), \dots, \nabla_{X_p} Y_i, \dots, Y_k(p), \omega(p)) + A(Y_1(p), \dots, Y_k(p), \nabla_{X_p} \omega). \end{aligned}$$

(b) If A is a tensor field of type $\binom{k}{l}$, where each $A(p)$ is regarded as a map $M_p \times \cdots \times M_p \times M_p^* \times \cdots \times M_p^* \rightarrow \mathbb{R}$, show that

$$\begin{aligned} (\nabla_{X_p} A)(Y_1(p), \dots, \omega_l(p)) &= \nabla_{X_p}(A(Y_1, \dots, \omega_l)) \\ &- \sum_{i=1}^k A(\dots, \nabla_{X_p} Y_i, \dots, \omega_l(p)) \\ &+ \sum_{i=1}^l A(Y_1(p), \dots, \nabla_{X_p} \omega_i, \dots). \end{aligned}$$

Consider in particular the cases $l = 0$ and $k = 0$.

(c) If we instead regard each $A(p)$ as a map from $M_p \times \cdots \times M_p$ to the set of maps $M_p^* \times \cdots \times M_p^* \rightarrow \mathbb{R}$, then

$$\begin{aligned} (\nabla_{X_p} A)(Y_1(p), \dots, Y_k(p)) &= \nabla_{X_p}(A(Y_1, \dots, Y_k)) \\ &- \sum_{i=1}^k A(Y_1(p), \dots, \nabla_{X_p} Y_i, \dots, Y_k(p)). \end{aligned}$$

2. Consider the situation on page 12. Writing

$$\frac{\partial}{\partial x^i} = \sum_{\rho} \frac{\partial y^{\rho}}{\partial x^i} \frac{\partial}{\partial y^{\rho}} = \sum_{\rho} y^{\rho}{}_{;i} \frac{\partial}{\partial y^{\rho}},$$

and similarly for $\partial/\partial x^j$, show that

$$\nabla'_{\partial/\partial x^j} \partial/\partial x^i = \sum_{\alpha} \left(\sum_{\rho, \sigma} \Gamma'_{\rho\sigma}{}^{\alpha} y^{\rho}{}_{;i} y^{\sigma}{}_{;j} \right) \frac{\partial}{\partial y^{\alpha}} + \sum_{\alpha} \left(\sum_{\rho} y^{\rho}{}_{;j} \frac{\partial y^{\alpha}{}_{;i}}{\partial y^{\rho}} \right) \frac{\partial}{\partial y^{\alpha}}.$$

Using

$$y^{\alpha}{}_{;ij} = \frac{\partial y^{\alpha}{}_{;i}}{\partial x^j} - \sum_k y^{\alpha}{}_{;k} \Gamma_{ji}^k = \sum_{\rho} y^{\rho}{}_{;j} \frac{\partial y^{\alpha}{}_{;i}}{\partial y^{\rho}} - \sum_k y^{\alpha}{}_{;k} \Gamma_{ji}^k,$$

verify the assertion near the bottom of page 13.

3. (The calculations in this Problem are similar to those on pages 79–82 and 97–100 and might be postponed until then, or they may be regarded as a rehearsal for the latter.) Let $\mathbf{X} = X_1, \dots, X_m$ and $\mathbf{X}' = X'_1, \dots, X'_m$ be two adopted orthonormal moving frames on $M^n \subset N$. Let s_{ij}^r and $s_{ij}'^r$ be the unique functions with

$$\psi_j^r = \sum_i s_{ij}^r \theta^i, \quad \psi_j'^r = \sum_i s_{ij}'^r \theta'^i.$$

Say that $\mathbf{X}' = \mathbf{X} \cdot a$ for an orthogonal matrix of functions a , so that we have (pp. II.280, 282)

$$\theta' = a^{-1} \cdot \theta \quad \text{and} \quad \psi' = a^{-1} da + a^{-1} \psi a.$$

The matrix a must satisfy $a_r^r = 0 = a_r^j$, since \mathbf{X} and \mathbf{X}' are both adopted to M . Conclude that

$$\psi_j'^r = \sum_{i,t} (a^{-1})_t^r \psi_i^t a_j^i,$$

and thus that

$$\sum_i s_{ij}'^r (a^{-1})_h^i = \sum_{i,t} (a^{-1})_t^r s_{ih}^t a_j^i \implies \sum_r s_{kj}'^r a_r^u = \sum_{i,h} s_{ih}^u a_j^i a_k^h.$$

Hence show that the definition

$$s(X'_j, X'_k) = \sum_r s_{jk}'^r X'_r$$

is compatible with the definition

$$s(X_i, X_h) = \sum_u s_{ih}^u X_u.$$

4. Apply the equation for ψ_j^{n+1} on page 20 to (X_k, X_l) . Noting that

$$\begin{aligned} d\psi_j^{n+1}(X_k, X_l) &= X_k(\psi_j^{n+1}(X_l)) - X_l(\psi_j^{n+1}(X_k)) - \psi_j^{n+1}([X_k, X_l]) \\ [X_k, X_l] &= \nabla'_{X_k} X_l - \nabla'_{X_l} X_k = \sum_i \omega_l^i(X_k) X_i - \omega_k^i(X_l) X_i, \end{aligned}$$

deduce the last equation in the proof of Theorem 11.

5. Let N and \bar{N} be two n -dimensional Riemannian manifolds of constant curvature K_0 . Let X_1, \dots, X_n be an orthonormal basis of N_p , and $\bar{X}_1, \dots, \bar{X}_n$ be an orthonormal basis of $\bar{N}_{\bar{p}}$. Let $c: [0, 1] \rightarrow N$ be a curve with $c(0) = p$. By Corollary II.7-13, there is an isometry f from a neighborhood U_0 of p to a neighborhood of \bar{p} , with $f_*(X_i) = \bar{X}_i$. A **continuation** of f along c is a family $\{f_t\}$ of isometries $f_t: U_t \rightarrow \bar{N}$ where U_t is a neighborhood of $c(t)$, with $f_0 = f$, satisfying the following condition: for each t there is $\delta > 0$ so that $|t - t'| < \delta \Rightarrow f_t = f_{t'}$ on $U_t \cap U_{t'}$.

(a) If $\{f_t\}$ and $\{g_t\}$ are two continuations of f , then each $f_t = g_t$ in some neighborhood of $c(t)$.

(b) If N is connected, then there is at most one isometry $\phi: N \rightarrow \bar{N}$ with $\phi_* X_i = \bar{X}_i$.

(c) Let \bar{N} be complete, and let $K = \{q \in \bar{N} : d(\bar{p}, q) \leq \text{length of } c\}$. Then K is compact (see pg. I.343). Conclude that there is $\delta > 0$ such that for any orthonormal $Y_1, \dots, Y_n \in N_{c(t)}$ ($0 \leq t \leq 1$) and orthonormal $\bar{Y}_1, \dots, \bar{Y}_n \in \bar{N}_q$ ($q \in K$), there is an isometry taking Y_i to \bar{Y}_i whose domain contains all $c(t')$ for $|t - t'| < \delta$. Then show that a continuation always exists.

(d) If N is simply-connected, show that for two paths $c, \gamma: [0, 1] \rightarrow M$ with $c(0) = \gamma(0) = p$ and $c(1) = \gamma(1) = q$, the continuations $\{f_t\}, \{g_t\}$ of f along c and γ must satisfy $f_1(q) = g_1(q)$. Conclude that for \bar{N} complete and N simply-connected, there is a (unique) isometry $\phi: N \rightarrow \bar{N}$ with $\phi_* X_i = \bar{X}_i$.

(e) Any two simply-connected complete manifolds of constant curvature K_0 are isometric, and there is an isometry taking any orthonormal basis of such a manifold to any other orthonormal basis.

CHAPTER 2

ELEMENTS OF THE THEORY OF SURFACES IN \mathbb{R}^3

This chapter will parallel as closely as possible the first chapter of Volume II. Our interest will now be directed away from the intrinsic geometry of surfaces, and toward those properties which describe the particular ways they are immersed in \mathbb{R}^3 .

We recall that in our study of curves, we defined certain quantities, the curvature κ and the torsion τ , which describe the local appearance of a curve in \mathbb{R}^3 . The curvature was first defined in an extremely geometric way, by taking limits of circles passing through three points of the curve. But it could be defined quite simply as $\kappa = |\mathbf{t}'|$, and exactly this approach was used to define τ . We then showed that these quantities actually describe the curve completely, up to Euclidean motions. We also investigated certain global properties connected with positive curvature. Finally, we showed how our investigations could be reformulated in terms of Lie groups, with Theorem I.10-18 playing a leading role, and then went on to investigate properties of curves invariant under a different group of motions of Euclidean space.

Our investigation of surfaces will proceed in just this order; however it will differ from the study of curves in one important respect. For curves we found that the definition and study of κ and τ [or of the affine curvature κ] was greatly simplified by considering curves parameterized only by arclength [or affine arclength]. But for surfaces there is no natural choice of a parameterization; to a certain extent this is responsible for the considerably greater complications one encounters in surface theory.

In Chapter II.3B we considered a submanifold $M \subset \mathbb{R}^3$, with $i: M \rightarrow \mathbb{R}^3$ the inclusion map, and we defined the **first fundamental form** I by $I = i^*\langle \ , \ \rangle$, where $\langle \ , \ \rangle$ is the usual Riemannian metric on \mathbb{R}^3 . In terms of a coordinate system $\chi = (x, y)$ on M , we wrote the tensor I on M as

$$I = E dx \otimes dx + F dx \otimes dy + F dy \otimes dx + G dy \otimes dy$$

for certain functions E, F, G on M ; and we noted that if the inverse of χ is

$f: U \rightarrow \mathbb{R}^3$ (for $U \subset \mathbb{R}^2$ open), then

$$\begin{aligned} E(f(s, t)) &= \left\langle \frac{\partial f}{\partial s}(s, t), \frac{\partial f}{\partial s}(s, t) \right\rangle = \langle f_1(s, t), f_1(s, t) \rangle \\ F(f(s, t)) &= \langle f_1(s, t), f_2(s, t) \rangle \\ G(f(s, t)) &= \langle f_2(s, t), f_2(s, t) \rangle. \end{aligned}$$

This means that $E = \langle f_1, f_1 \rangle \circ f^{-1}$, etc., which sometimes makes the functions E, F, G rather awkward to work with; consequently, we will often find it convenient to change our view slightly, and define everything explicitly in terms of a given immersion.

If $f: M \rightarrow \mathbb{R}^3$ is an immersion, we define the **first fundamental form** I_f of f to be the tensor $f^*\langle \ , \ \rangle$ on M . In particular, when $f: U \rightarrow \mathbb{R}^3$ (for $U \subset \mathbb{R}^2$ open) we have a form I_f on U , with

$$I_f(s, t)(v, w) = \langle f_*v, f_*w \rangle \quad \text{for } v, w \in \mathbb{R}^2_{(s, t)}.$$

We can then define functions E, F, G directly on U by

$$\begin{aligned} E &= \left\langle \frac{\partial f}{\partial s}, \frac{\partial f}{\partial s} \right\rangle = \langle f_1, f_1 \rangle \\ F &= \langle f_1, f_2 \rangle \\ G &= \langle f_2, f_2 \rangle. \end{aligned}$$

These functions are nothing but the components of $I_f = f^*\langle \ , \ \rangle$ with respect to the standard coordinate system (s, t) on \mathbb{R}^2 [and they have essentially already been introduced on pg. II.128]. Since $f^*\langle \ , \ \rangle$ is positive definite, we have $EG - F^2 > 0$ (pg. I.308); moreover (see Problem I.9-5), we have

$$|f_1 \times f_2| = \sqrt{EG - F^2}.$$

It will often be much more convenient to use the subscript notation, which was classically used for the higher dimensional cases,

$$g_{ij} = \langle f_i, f_j \rangle, \quad \text{so that} \quad g_{11} = E, \quad g_{12} = g_{21} = F, \quad g_{22} = G.$$

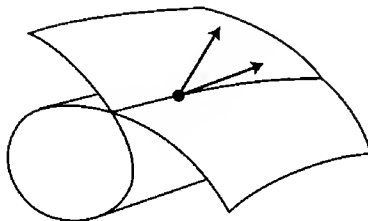
We also introduce the functions g^{ij} which satisfy

$$\sum_k g_{ik} g^{kj} = \delta_i^j.$$

The functions E, F, G are analogous to the single function $t \mapsto |c'(t)|$, defined for a curve c . We usually reparameterized our curves so that this function was equal to 1, but for surfaces, where no convenient reparameterization is available, the functions E, F, G always play a vital role, analogous to the arclength function of a curve.

We next seek an analogue of the functions κ and τ of a curve c . The definition of κ and τ depended very much on the possibility of parameterizing c by arclength, so that $\mathbf{t}(s) = c'(s)$ has length 1, and consequently $\mathbf{t}'(s) = \kappa(s)\mathbf{n}(s)$ for some unit vector $\mathbf{n}(s)$ perpendicular to the curve. In the case of a surface $M \subset \mathbb{R}^3$, we do not have a special parameterization to work with, but we already have an analogue of \mathbf{n} , namely a unit normal field ν , which can at least be defined in a neighborhood of each point. We recall that a choice of ν is equivalent to a choice of an orientation for M , for we can let $(X_1, X_2) \in M_p$ be positively oriented if and only if $(X_1, X_2, \nu(p))$ is positively oriented in \mathbb{R}^3 .

When we are not dealing with a submanifold, but with an immersion $f: M \rightarrow \mathbb{R}^3$, the normal field should be considered as a “vector field along f ”, since we may have points $p, q \in M$ with $f(p) = f(q)$, but with different normals at this



point. We will denote this vector field along f by

$$q \mapsto N(q)_{f(q)} \in \mathbb{R}^3_{f(q)}.$$

Thus N is a function $N: M \rightarrow S^2 \subset \mathbb{R}^3$. We will always adhere to the convention of using ν when we are specifically considering imbedded submanifolds $M \subset \mathbb{R}^3$, and using N when we are considering immersions (and imbeddings) $f: M \rightarrow \mathbb{R}^3$; when necessary, we will write N_f to indicate the dependence of N on f . If $W \subset M$ is an open set on which f is an imbedding, then a unit normal field ν on $M = f(W) \subset \mathbb{R}^3$ is determined by the condition that $N = \nu \circ f$ on W (naturally we have to regard ν as a map $\nu: M \rightarrow S^2 \subset \mathbb{R}^3$ in order to write this). In terms of ν we have already defined the **second fundamental form II on M** by

$$\begin{aligned} \text{II}(p)(v_p, w_p) &= \langle -d\nu(v_p), w_p \rangle \\ &= \langle -\nu_*(v_p), w_p \rangle \quad \text{for } v_p, w_p \in M_p. \end{aligned}$$

We can define the **second fundamental form** \mathbf{II}_f of f to be the tensor on M defined by

$$\begin{aligned}\mathbf{II}_f(q)(v_q, w_q) &= \langle -dN(v_q), f_*(w_q) \rangle \\ &= \langle -N_*(v_q), f_*(w_q) \rangle \quad \text{for } v_q, w_q \in M_q.\end{aligned}$$

Equivalently, we have $\mathbf{II}_f = f^*\mathbf{II}$.

In particular, let us consider an immersion $f: U \rightarrow \mathbb{R}^3$, for $U \subset \mathbb{R}^2$ open. We can specifically choose N to be

$$N = \frac{f_1 \times f_2}{|f_1 \times f_2|} = \frac{f_1 \times f_2}{\sqrt{EG - F^2}} \quad \begin{array}{l} \text{(the negative square root gives} \\ \text{the other possible choice for } N\text{).} \end{array}$$

Then

$$\begin{aligned}\mathbf{II}_f(s, t)(v, w) &= \langle -dN(v), f_*(w) \rangle \\ &= \langle -N_*(v), f_*(w) \rangle \quad v, w \in \mathbb{R}^2_{(s, t)}.\end{aligned}$$

We will define functions l, m, n directly on U by

$$\begin{aligned}l &= \left\langle -\frac{\partial N}{\partial s}, \frac{\partial f}{\partial s} \right\rangle = \langle -N_1, f_1 \rangle = \langle N, f_{11} \rangle \\ m &= \langle -N_1, f_2 \rangle = \langle N, f_{12} \rangle \\ n &= \langle -N_2, f_2 \rangle = \langle N, f_{22} \rangle.\end{aligned}$$

These functions l, m, n are simply the components of \mathbf{II}_f with respect to the standard coordinate system (s, t) on \mathbb{R}^2 (compare with the proof of Theorem 1 on pg. II.123). Once again, it will often be more convenient to use subscript notation:

$$l_{ij} = \langle -N_i, f_j \rangle = \langle N, f_{ij} \rangle.$$

Before we go any further, we should clear up one problem. Let us suppose that $f: U \rightarrow \mathbb{R}^3$ is actually an imbedding, and let $M = f(U) \subset \mathbb{R}^3$. In geometric considerations, it is usually the linear transformation $-dv: M_p \rightarrow M_p$ which interests us, rather than the second fundamental form \mathbf{II} itself. Now for $p = f(s, t)$, the map $-dv: M_p \rightarrow M_p$ is related to the matrix $(l_{ij}(s, t))$ in the following way:

$$l_{ij}(s, t) = \langle -dv(f_i(s, t)_p), f_j(s, t)_p \rangle.$$

Unfortunately, this does *not* mean that $(l_{ij}(s, t))$ is the matrix of $-dv$ with respect to the basis $f_1(s, t)_p, f_2(s, t)_p$, because these vectors are not necessarily orthonormal. To avoid going out of our minds now, and especially in the next chapter, we make note of the following:

0. FACT. Let v_1, \dots, v_n be a basis for the vector space V , with the inner product $\langle \cdot, \cdot \rangle$. Suppose that A is the matrix of a linear transformation $T: V \rightarrow V$ with respect to v_1, \dots, v_n , while B and C are the matrices $B = (\langle T v_i, v_j \rangle)$ and $C = (\langle v_i, v_j \rangle)$. Then

$$A = (BC^{-1})^t,$$

where t denotes the transpose.

PROOF. We have $T v_i = \sum_j A_{ji} v_j$, and consequently

$$B_{ik} = \langle T v_i, v_k \rangle = \sum_j A_{ji} \langle v_j, v_k \rangle = (A^t \cdot C)_{ik},$$

so $B = A^t \cdot C$. ♦

We apply this observation with

$$A = \begin{pmatrix} \text{matrix of } -dv: M_p \rightarrow M_p \\ \text{with respect to } (f_1)_p, (f_2)_p \end{pmatrix}$$

and

$$B = (l_{ij}), \quad C = (g_{ij}),$$

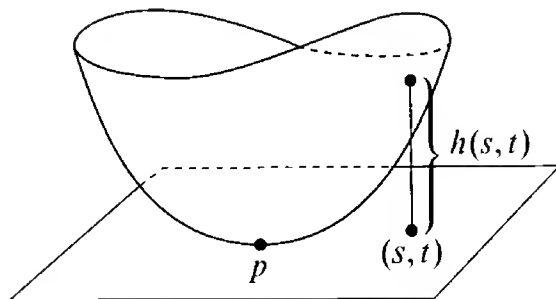
where f_i, l_{ij}, g_{ij} are all evaluated at (s, t) , and $p = f(s, t)$. Since B and C are symmetric, we have $(BC^{-1})^t = C^{-1}B$; so we find that

$$\begin{aligned} \text{(I)} \quad \begin{pmatrix} \text{matrix of } -dv: M_p \rightarrow M_p \\ \text{with respect to } (f_1)_p, (f_2)_p \end{pmatrix} &= (g_{ij})^{-1}(l_{ij}) \\ &= \frac{1}{EG - F^2} \cdot \begin{pmatrix} G & -F \\ -F & E \end{pmatrix} \begin{pmatrix} l & m \\ m & n \end{pmatrix} \\ &[f_i, g_{ij}, l_{ij} \text{ evaluated at } (s, t); p = f(s, t)]. \end{aligned}$$

The functions l, m, n certainly seem to be good candidates for the desired analogues of the functions κ and τ . As a first test of their appropriateness, we will investigate how well these functions describe f up to second order in a neighborhood of a point p .

What we are interested in is the shape of $M = \text{image } f$, not the particular parameterization f itself; so we will essentially fix the parameterization by describing our surface M in terms of the distance of a point from the tangent plane at p . For convenience we will assume that $p = 0 \in \mathbb{R}^3$ and that the

tangent plane at p is the (x, y) -plane. Then our surface M is the graph of a



function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ with $h_1(0, 0) = h_2(0, 0) = 0$. Applying Taylor's formula for functions of two variables, we have

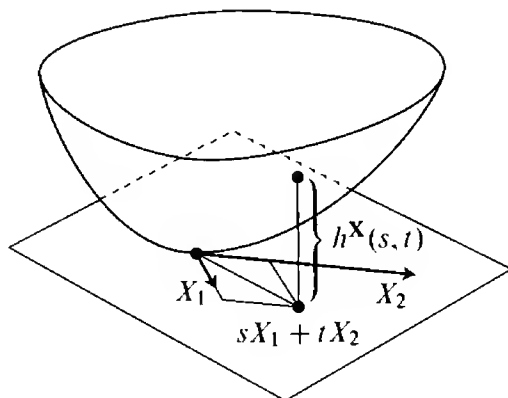
$$h(s, t) = \frac{1}{2}(h_{11}(0, 0) \cdot s^2 + 2h_{12}(0, 0) \cdot st + h_{22}(0, 0) \cdot t^2) + R(s, t),$$

where $R(s, t)/|(s, t)|^2 = R(s, t)/(s^2 + t^2) \rightarrow 0$ as $(s, t) \rightarrow 0$. We therefore say that the quadratic surface

$$P = \left\{ (s, t, \frac{1}{2}(h_{11}(0, 0) \cdot s^2 + 2h_{12}(0, 0) \cdot st + h_{22}(0, 0) \cdot t^2)) \right\}$$

“approximates M up to order 2 at 0”.

Now we want to make an elementary observation which is crucial for avoiding confusion. Suppose that $\mathbf{X} = (X_1, X_2)$ is any basis for \mathbb{R}^2 , say with $X_1 = (a_{11}, a_{21})$ and $X_2 = (a_{12}, a_{22})$. Our surface M can also be described as the graph of a function in terms of the “ $X_1, X_2, (0, 0, 1)$ coordinate system”. In



other words, we can consider the function $h^{\mathbf{X}}$ with $h^{\mathbf{X}}(s, t) =$ the height above the (x, y) -plane of the point of M lying above $sX_1 + tX_2$. This means that

$$h^{\mathbf{X}}(s, t) = h(sX_1 + tX_2) = h(a_{11}s + a_{12}t, a_{21}s + a_{22}t).$$

If we momentarily denote (s, t) by (s^1, s^2) , then we can write more conveniently

$$h^{\mathbf{X}}(s^1, s^2) = h\left(\sum_{i=1}^2 a_{1i}s^i, \sum_{i=1}^2 a_{2i}s^i\right),$$

and for the partial derivatives of $h^{\mathbf{X}}$ we easily compute that

$$\begin{aligned} h^{\mathbf{X}}_{\alpha}(s^1, s^2) &= \sum_{j=1}^2 a_{j\alpha} h_j\left(\sum_{i=1}^2 a_{1i}s^i, \sum_{i=1}^2 a_{2i}s^i\right) \\ h^{\mathbf{X}}_{\alpha\beta}(s^1, s^2) &= \sum_{j,k=1}^2 a_{j\alpha} a_{k\beta} h_{jk}\left(\sum_{i=1}^2 a_{1i}s^i, \sum_{i=1}^2 a_{2i}s^i\right), \end{aligned}$$

so that in particular

$$(*) \quad \begin{cases} h^{\mathbf{X}}_{\alpha}(0, 0) = 0 \\ h^{\mathbf{X}}_{\alpha\beta}(0, 0) = \sum_{j,k=1}^2 a_{j\alpha} a_{k\beta} h_{jk}(0, 0). \end{cases}$$

Now the quadratic surface $Q \subset \mathbb{R}^3 = \mathbb{R}^2 \times \mathbb{R}$ defined by

$$Q = \left\{ (sX_1 + tX_2, \frac{1}{2}(h^{\mathbf{X}}_{11}(0, 0) \cdot s^2 + 2h^{\mathbf{X}}_{12}(0, 0) \cdot st + h^{\mathbf{X}}_{22}(0, 0) \cdot t^2)) \right\}$$

can equally well be said to approximate M up to order 2 at 0. But we claim that the surfaces P and Q are *exactly the same*. The best way to express this claim is as follows. For each basis $\mathbf{X} = (X_1, X_2)$, let us define a function $\Phi^{\mathbf{X}}: \mathbb{R}^2 \rightarrow \mathbb{R}$ by

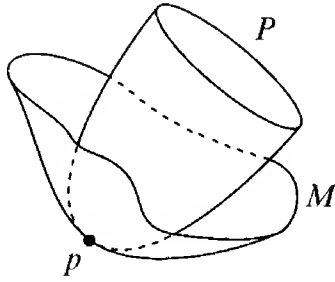
$$\Phi^{\mathbf{X}}(sX_1 + tX_2) = \frac{1}{2}(h^{\mathbf{X}}_{11}(0, 0) \cdot s^2 + 2h^{\mathbf{X}}_{12}(0, 0) \cdot st + h^{\mathbf{X}}_{22}(0, 0) \cdot t^2).$$

Then the functions $\Phi^{\mathbf{X}}$ are all the *same* function $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}$, and we can therefore describe both P and Q simply as the graph of Φ . To check that all $\Phi^{\mathbf{X}}$ are the same, let us use Φ for the function we obtain when \mathbf{X} is the standard basis $(1, 0), (0, 1)$. Then

$$\begin{aligned} \Phi(s^1 X_1 + s^2 X_2) &= \Phi\left(\sum_{i=1}^2 a_{1i}s^i, \sum_{i=1}^2 a_{2i}s^i\right) \\ &= \frac{1}{2} \sum_{j,k=1}^2 h_{jk}(0, 0) \cdot \left(\sum_{i=1}^2 a_{ji}s^i\right) \left(\sum_{i=1}^2 a_{ki}s^i\right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{\alpha, \beta=1}^2 \left(\sum_{j, k=1}^2 h_{jk}(0, 0) a_{j\alpha} a_{k\beta} \right) s^\alpha s^\beta \\
&= \frac{1}{2} \sum_{\alpha, \beta=1}^2 h^{\mathbf{X}}_{\alpha\beta}(0, 0) s^\alpha s^\beta \quad \text{by } (*) \\
&= \Phi^{\mathbf{X}}(s^1 X_1 + s^2 X_2),
\end{aligned}$$

which is what we wanted to prove. It is also easy to see that if we describe M in terms of the “ $X_1, X_2, (0, 0, -1)$ coordinate system”, then Φ becomes $-\Phi$, while the resulting second order approximating surface is unchanged. Thus, for every point p of a surface M in \mathbb{R}^3 , there is a well-defined quadratic surface P which approximates M up to order 2 at p .



Let us for simplicity stick to the case where $p = 0 \in \mathbb{R}^3$, the tangent plane at p is the (x, y) -plane, and M is the graph of $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ (in the standard coordinate system). The surface M is the image of the immersion

$$f(s, t) = (s, t, h(s, t))$$

for which we have

$$\begin{aligned}
N &= N(0, 0) = (0, 0, 1) \\
l &= l(0, 0) = \langle (0, 0, 1), (0, 0, h_{11}(0, 0)) \rangle \\
&= h_{11}(0, 0) \\
m &= m(0, 0) = h_{12}(0, 0) \\
n &= n(0, 0) = h_{22}(0, 0).
\end{aligned}$$

Thus our approximating quadratic surface P at 0 is described explicitly as the graph of

$$\alpha(s, t) = \frac{1}{2}(ls^2 + 2mst + nt^2) = \frac{1}{2} \left\langle (s, t), (s, t) \cdot \begin{pmatrix} l & m \\ m & n \end{pmatrix} \right\rangle.$$

To see just what this graph looks like, we choose two *orthonormal* eigenvectors $X_1, X_2 \in \mathbb{R}^2$ for the symmetric matrix $\begin{pmatrix} l & m \\ m & n \end{pmatrix}$, with corresponding eigenvalues k_1, k_2 . Then

$$\begin{aligned} \alpha(sX_1 + tX_2) &= \frac{1}{2} \left\langle sX_1 + tX_2, sX_1 + tX_2 \cdot \begin{pmatrix} l & m \\ m & n \end{pmatrix} \right\rangle \\ &= \frac{1}{2} \langle sX_1 + tX_2, sk_1X_1 + tk_2X_2 \rangle \\ &= \frac{1}{2} (k_1s^2 + k_2t^2). \end{aligned}$$

In other words, after a rotation of our axes so that they point along X_1, X_2 , the graph of α becomes the graph of $\tilde{\alpha}: \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

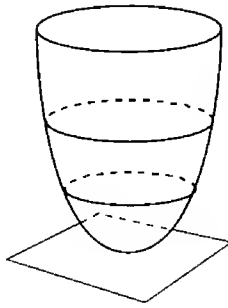
$$\tilde{\alpha}(s, t) = \frac{1}{2} (k_1s^2 + k_2t^2).$$

The shape of this graph depends on the sign of $k_1k_2 = \det \begin{pmatrix} l & m \\ m & n \end{pmatrix} = ln - m^2$, and leads us to classify the points p into four types.

1. *Elliptic point*: $ln - m^2 > 0$. Then k_1, k_2 have the same sign. If $k_1, k_2 > 0$, then the graph of

$$\tilde{\alpha}(s, t) = \frac{s^2}{(\sqrt{2/k_1})^2} + \frac{t^2}{(\sqrt{2/k_2})^2}$$

is an *elliptic paraboloid*; planes parallel to the (x, y) -plane intersect the graph of $\tilde{\alpha}$ in similar ellipses, while planes parallel to the other coordinate planes intersect the graph in parabolas.



In our new coordinate system, the original surface is the graph of \tilde{h} , where

$$\tilde{h}(s, t) = \tilde{\alpha}(s, t) + \tilde{R}(s, t).$$

with $\tilde{R}(s, t)/(s^2 + t^2) \rightarrow 0$. There is clearly a constant $A > 0$ such that $\tilde{\alpha}(s, t) > A(s^2 + t^2)$, so

$$0 = \lim_{(s,t) \rightarrow 0} -\frac{\tilde{R}(s, t)}{s^2 + t^2} = \lim_{(s,t) \rightarrow 0} \left(\frac{\tilde{\alpha}(s, t) - \tilde{h}(s, t)}{s^2 + t^2} \right) \geq A - \lim_{(s,t) \rightarrow 0} \frac{\tilde{h}(s, t)}{s^2 + t^2},$$

and hence

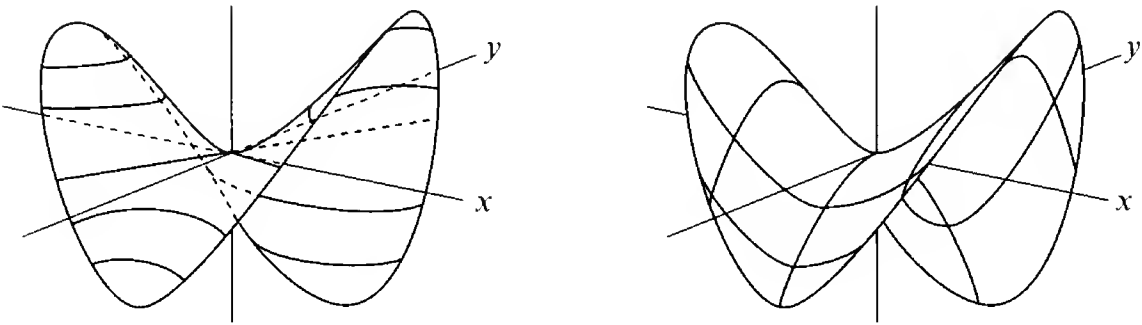
$$\frac{\tilde{h}(s, t)}{s^2 + t^2} \geq A/2 > 0 \quad \text{for sufficiently small } (s, t).$$

Therefore $\tilde{h}(s, t) > 0$ for small (s, t) . Thus points of our surface which are near p lie on the same side of the tangent plane at p as $v(p)$. If $k_1, k_2 < 0$, then the graph of $\tilde{\alpha}$ is an elliptic paraboloid pointing in the other direction, and points of our surface which are near p lie on the other side of the tangent plane at p .

2. *Hyperbolic point:* $ln - m^2 < 0$. Then k_1, k_2 have opposite signs, say $k_1 > 0 > k_2$. The graph of

$$\tilde{\alpha}(s, t) = \frac{s^2}{(\sqrt{2/k_1})^2} - \frac{t^2}{(\sqrt{-2/k_2})^2}$$

is a *hyperbolic paraboloid*; planes parallel to the (x, y) -plane intersect the graph of $\tilde{\alpha}$ in similar hyperbolas [except that the (x, y) -plane itself intersects the graph in two straight lines through $(0, 0)$], while planes parallel to the other coordinate planes intersect the graph in parabolas. It is easy to see that there are points

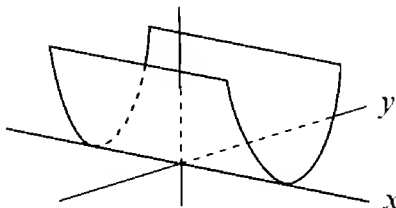


of the original surface arbitrarily close to p lying on both sides of the tangent plane at p .

3. *Parabolic point:* $ln - m^2 = 0$, but not all of l, m, n are 0. Then exactly one eigenvalue is 0; say $k_1 = 0$ but $k_2 \neq 0$. The graph of

$$\tilde{\alpha}(s, t) = \frac{1}{2}k_2 t^2$$

is a *parabolic cylinder*. If $k_2 > 0$, then our original surface must contain points close to p on the same side of the tangent plane as $v(p)$. But there can also be



points arbitrarily close to p on the other side of the tangent plane. For example, our surface might be the graph of

$$h(s, t) = s^3 + t^2.$$

4. *Planar point*: $l = m = n = 0$. The graph of $\tilde{\alpha}$ is the (x, y) -plane.

In the planar case, nothing at all can be said about which side of the tangent plane our surface lies on. For example, our surface might be the graph of any one of the following functions:

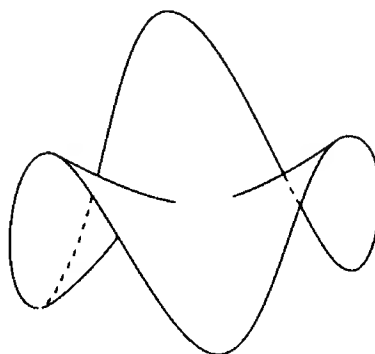
$$h(s, t) = s^4 \quad \text{graph lies above the } (x, y)\text{-plane}$$

$$h(s, t) = -s^4 \quad \text{graph lies below the } (x, y)\text{-plane}$$

$$h(s, t) = s^3 \quad \text{graph lies above and below the } (x, y)\text{-plane.}$$

A more interesting example of a planar point is provided by the “monkey saddle”, the graph of

$$\begin{aligned} h(s, t) &= s^3 - 3st^2 \\ &= \text{real part of } (s + it)^3, \end{aligned}$$

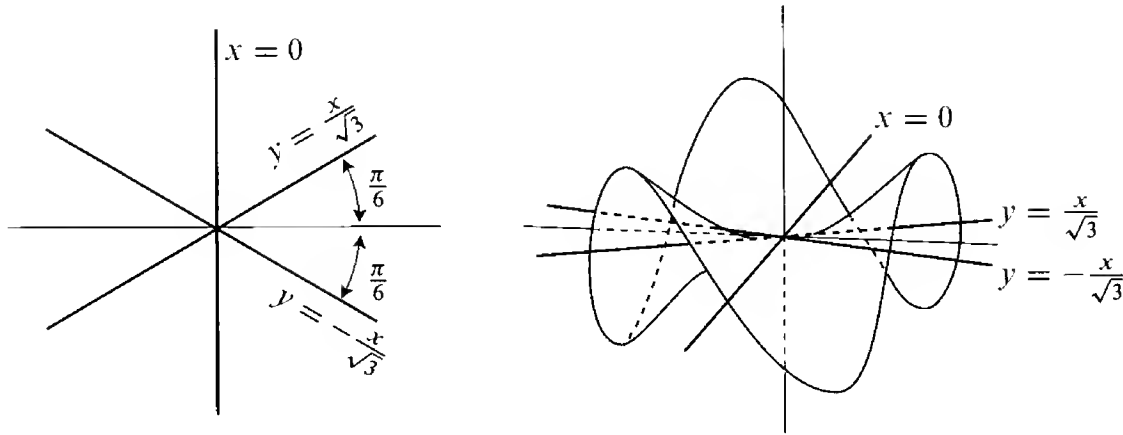


with a planar point at $0 \in \mathbb{R}^3$. I was always very confused by the name of this surface, because I thought it was supposed to be a saddle that you put on a monkey. Actually it's a saddle that a monkey *uses* (to ride a bicycle, say)—there are two depressions for its legs, and an extra one for its tail. The monkey saddle

intersects the (x, y) -plane in the set

$$\{(x, y) : x^3 - 3xy^2 = 0\}$$

which consists of 3 straight lines all making equal angles with each other.



Notice that our classification of points on a surface as elliptic, hyperbolic, parabolic, or planar, does not at all depend on the special parameterization which we introduced; for equation (I) on page 35 shows that

$$\det(-dv : M_p \rightarrow M_p) = \frac{\det(l_{ij})}{\det(g_{ij})} = \frac{ln - m^2}{EG - F^2},$$

which means that the sign of $ln - m^2$ is always the same as the sign of $\det(-dv)$. When we do introduce our special parameterization, we have $E = G = 1$ and $F = 0$ at $(0, 0)$, so equation (I) then gives

$$(l_{ij}) = \text{matrix of } -dv : M_p \rightarrow M_p.$$

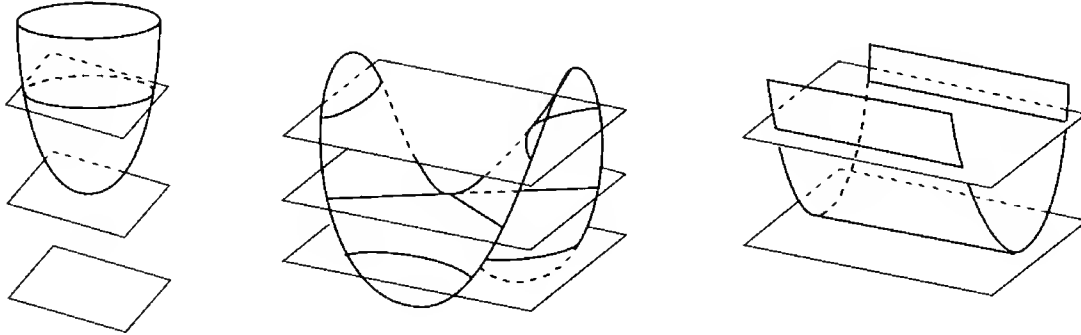
Consequently, the numbers k_1, k_2 which we have found can also be described invariantly as the eigenvalues of $-dv$; the orthonormal vectors X_1, X_2 (in \mathbb{R}^2 , which we have identified with M_p) are just the eigenvectors of $-dv$.

The quadratic surface P which approximates M up to order 2 at a point $p \in M$ is called the **osculating paraboloid** at p ; when $p = 0 \in \mathbb{R}^2$, the tangent plane at p is the (x, y) -plane, and X_1, X_2 point along the x - and y -axes, it is the graph of

$$\alpha(s, t) = \tilde{\alpha}(s, t) = \frac{1}{2}(k_1 s^2 + k_2 t^2).$$

For space curves we obtained an analogous osculating curve (pg. II.31), and we used this osculating curve to examine the original curve more closely by projecting it on the coordinate planes. That procedure wouldn't make much sense here, but there is something else we can do. Suppose we first intersect

the osculating paraboloid with the two planes parallel to the (x, y) -plane and at distance d from it, and then project the intersection onto the (x, y) -plane.



We obtain the set

$$J_d = \{(x, y) : k_1 x^2 + k_2 y^2 = \pm 2d\},$$

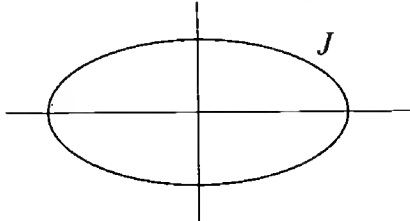
which is either an ellipse, a pair of hyperbolas, a pair of parallel lines, or nothing. Clearly, the sets

$$\frac{J_d}{\sqrt{2d}} = \left\{ \left(\frac{x}{\sqrt{2d}}, \frac{y}{\sqrt{2d}} \right) : (x, y) \in J_d \right\}$$

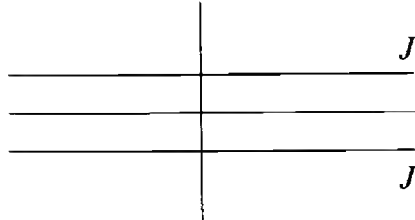
are all the same, namely

$$\frac{J_d}{\sqrt{2d}} = J = \{(x, y) : k_1 x^2 + k_2 y^2 = \pm 1\}.$$

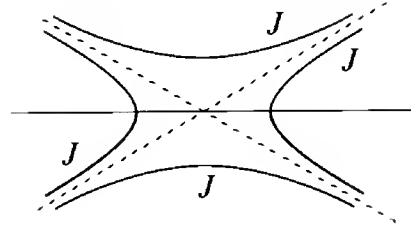
(a) k_1, k_2 have same sign



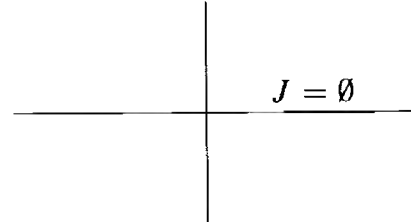
(c) $k_1 = 0, k_2 \neq 0$



(b) k_1, k_2 have opposite signs



(d) $k_1 = k_2 = 0$



Suppose now that we repeat this procedure, except that we intersect the two planes with our original surface instead of with its osculating paraboloid. We would expect the limiting set to be the same, since the surface agrees with its osculating paraboloid up to order 2. Actually, one has to be a little careful in formulating this result, and the corresponding proof is somewhat long, but not very interesting.

1. PROPOSITION. Let $p \in M$ be a point of an imbedded surface $M \subset \mathbb{R}^3$, and let X_1, X_2 be orthonormal eigenvectors of $-dv: M_p \rightarrow M_p$, with corresponding eigenvalues k_1, k_2 . Let $I \subset M_p$ be the set

$$I = \{X \in M_p : k_1 \langle X, X_1 \rangle^2 + k_2 \langle X, X_2 \rangle^2 = \pm 1\},$$

so that I is congruent to

- (a) the ellipse $k_1 x^2 + k_2 y^2 = \pm 1$ p an elliptic point
- (b) the hyperbolas $k_1 x^2 + k_2 y^2 = \pm 1$ p a hyperbolic point
- (c) the parallel lines $k_2 y^2 = \pm 1$ p a parabolic point with $k_2 \neq 0$
- (d) \emptyset p a planar point.

For $d > 0$, let I_d be the projection on M_p of the intersection of M with the two planes parallel to M_p and at distance d from it. Then

$$\lim_{d \rightarrow 0} \frac{I_d}{\sqrt{2d}} = I,$$

where this limit has the following meaning: For every $\varepsilon > 0$, and every compact set $C \subset M_p$, there is some $\delta > 0$ such that if $0 < d < \delta$, then

- (i) every point of $I \cap C$ is within ε of some point of $I_d/\sqrt{2d}$
- (ii) every point of $(I_d/\sqrt{2d}) \cap C$ is within ε of some point of I .

Remark: In case (d), this just means that $I_d/\sqrt{2d}$ eventually lies outside of any compact set. In case (a), we clearly do not have to use the compact set C in condition (i), since I itself is compact. And in condition (ii) the compact set C is needed only to exclude points of I_d coming from extraneous points of M which are not near p .

PROOF. We assume that $p = 0 \in \mathbb{R}^3$, and that M_p is the (x, y) -plane, with the eigenvectors X_1, X_2 of $-dv(p)$ pointing along the x - and y -axes. Then M is locally the graph of a function $h: U \rightarrow \mathbb{R}$ (for $U \subset \mathbb{R}^2$ open), and we can assume that

$$I_d = \{(s, t) \in U : h(s, t) = \pm d\}$$

[there is no need to consider the points (s, t) outside U , since for sufficiently small d , the corresponding points $(s/\sqrt{2d}, t/\sqrt{2d})$ lie outside of any given compact set $C \subset M_p$]. We thus have

$$I_d = \left\{ (s, t) \in U : \frac{k_1 s^2}{2} + \frac{k_2 t^2}{2} + R(s, t) = \pm d \right\},$$

where

$$(1) \quad \frac{R(s, t)}{s^2 + t^2} \rightarrow 0 \quad \text{as } (s, t) \rightarrow 0.$$

Hence

$$\frac{I_d}{\sqrt{2d}} = \left\{ (\sigma, \tau) : \begin{array}{l} (\sqrt{2d}\sigma, \sqrt{2d}\tau) \in U \text{ and} \\ k_1\sigma^2 + k_2\tau^2 + \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} = \pm 1 \end{array} \right\}.$$

Setting $s = \sqrt{2d}\sigma$ and $t = \sqrt{2d}\tau$ in (1), we see that

$$(2) \quad \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d(\sigma^2 + \tau^2)} \rightarrow 0 \quad \text{as } (\sqrt{2d}\sigma, \sqrt{2d}\tau) \rightarrow 0.$$

Now suppose we are given $\varepsilon > 0$, and a compact set $C \subset M_p$, which we might as well assume is of the form

$$C = \{(\sigma, \tau) : \sqrt{\sigma^2 + \tau^2} \leq A\}.$$

Choose $\varepsilon_0 > 0$ so that

$$(3) \quad |\alpha| < \varepsilon_0 \implies |1 - \sqrt{1 \pm \alpha}| < \varepsilon/A.$$

Let $(\sigma, \tau) \in C \cap I$, so that

$$(4) \quad k_1\sigma^2 + k_2\tau^2 = \pm 1 \quad \text{and} \quad \sqrt{\sigma^2 + \tau^2} \leq A.$$

Consider the function

$$(5) \quad d \mapsto \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} \quad 0 < d \leq 1.$$

This is continuous in d , and approaches 0 as $d \rightarrow 0^+$, by (2). So there is $\delta > 0$ such that

$$(6) \quad 0 < d < \delta \implies \left| \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} \right| < \varepsilon_0.$$

Let

$$\alpha = \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d},$$

and consider the point

$$(\sigma\sqrt{1\mp\alpha}, \tau\sqrt{1\mp\alpha}) = \begin{cases} (\sigma\sqrt{1-\alpha}, \tau\sqrt{1-\alpha}) & \text{if } +1 \text{ holds in equation (4)} \\ (\sigma\sqrt{1+\alpha}, \tau\sqrt{1+\alpha}) & \text{if } -1 \text{ holds in equation (4)}. \end{cases}$$

We have

$$\begin{aligned} k_1(\sigma\sqrt{1-\alpha})^2 + k_2(\tau\sqrt{1-\alpha})^2 &= (1-\alpha)[k_1\sigma^2 + k_2\tau^2] = 1-\alpha \\ k_1(\sigma\sqrt{1+\alpha})^2 + k_2(\tau\sqrt{1+\alpha})^2 &= (1+\alpha)[k_1\sigma^2 + k_2\tau^2] = -1-\alpha, \end{aligned}$$

so $(\sigma\sqrt{1\mp\alpha}, \tau\sqrt{1\mp\alpha})$ is in $I_d/\sqrt{2d}$. Its distance from (σ, τ) is

$$\begin{aligned} |1 - \sqrt{1\mp\alpha}| \sqrt{\sigma^2 + \tau^2} &< \frac{\varepsilon}{A} \sqrt{\sigma^2 + \tau^2} && \text{by (3), since } |\alpha| < \varepsilon_0 \quad \text{by (6)} \\ &< \varepsilon && \text{by (4).} \end{aligned}$$

We have thus found a $\delta > 0$ so that the given point $(\sigma, \tau) \in C \cap I$ has distance less than ε from a point of $I_d/\sqrt{2d}$, for all $d < \delta$. To conclude that one δ can be found which works for all $(\sigma, \tau) \in C \cap I$ we need the fact that the function (5) approaches 0 uniformly in (σ, τ) ; this follows from compactness of C .

We will now prove (ii). Given $\varepsilon > 0$ and $A > 0$, pick $\varepsilon_0 > 0$ so that

$$(7) \quad |\alpha| < \varepsilon_0 \implies \left| 1 - \frac{1}{\sqrt{1\pm\alpha}} \right| < \frac{\varepsilon}{A}.$$

Then pick $\delta_0 > 0$ so that

$$s^2 + t^2 < \delta_0 \implies \frac{R(s, t)}{s^2 + t^2} < \frac{\varepsilon_0}{2A^2}.$$

Setting $s = \sqrt{2d}\sigma$ and $t = \sqrt{2d}\tau$, we see that

$$(8) \quad \sigma^2 + \tau^2 < \frac{\delta_0}{2d} \implies \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} < \frac{\varepsilon_0}{A^2}(\sigma^2 + \tau^2).$$

Let $\delta = \delta_0/2A^2$. Then

$$0 < d < \delta \implies \frac{\delta_0}{2d} > A^2.$$

So if $0 < d < \delta$, then

$$(9) \quad \text{either } \sigma^2 + \tau^2 > A^2 \quad \text{or} \quad \sigma^2 + \tau^2 < A^2 < \frac{\delta_0}{2d}.$$

In the first case, the point $(\sigma, \tau) \in I_d/\sqrt{2d}$ is at distance $> A$ from the origin. In the second case, we have

$$\begin{aligned} \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} &< \frac{\varepsilon_0}{A^2}(\sigma^2 + \tau^2) && \text{by (8)} \\ &< \varepsilon_0 && \text{by (9),} \end{aligned}$$

so the point $(\sigma, \tau) \in I_d/\sqrt{2d}$ satisfies the equation

$$\begin{aligned} (10) \quad k_1\sigma^2 + k_2\tau^2 &= \pm 1 - \frac{R(\sqrt{2d}\sigma, \sqrt{2d}\tau)}{d} \\ &= \pm 1 - \alpha, \quad \text{where } 0 \leq |\alpha| < \varepsilon_0. \end{aligned}$$

Now the point

$$\left(\frac{\sigma}{\sqrt{1 \mp \alpha}}, \frac{\tau}{\sqrt{1 \mp \alpha}} \right)$$

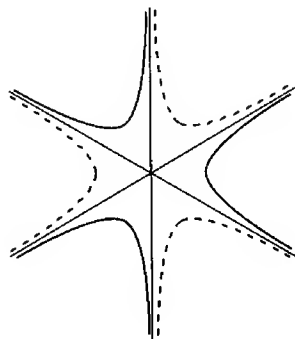
is in I , and its distance from (σ, τ) is

$$\begin{aligned} \left| 1 - \frac{1}{\sqrt{1 \mp \alpha}} \right| \sqrt{\sigma^2 + \tau^2} &< \frac{\varepsilon}{A} \sqrt{\sigma^2 + \tau^2} && \text{by (7), since } |\alpha| < \varepsilon_0 \text{ by (10)} \\ &< \varepsilon && \text{by (9).} \end{aligned}$$

This completes the proof. \blacklozenge

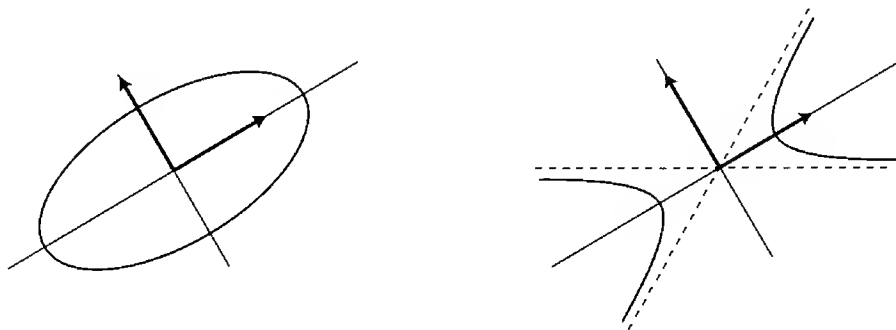
The limiting set $I \subset M_p$ of Proposition 1 is called the **Dupin indicatrix** at p . In the case of a planar point, we can obtain a more meaningful indicatrix by considering $\lim_{d \rightarrow 0} I_d/\sqrt[3]{d}$ —the adjustment factor $1/\sqrt[3]{d}$ is just what we need in

order that the projection on M_p of the intersections of parallel planes with the osculating *cubic* will be the same. The figure below shows the resulting indicatrix for the monkey saddle; the continuous lines come from intersections with planes



on one side of the tangent plane, and the dashed lines from intersections with planes on the other side. Similarly, if all derivatives of $h: U \rightarrow \mathbb{R}$ up to order $k - 1$ are 0 at $(0, 0)$, then we can look at the generalized indicatrix $\lim_{d \rightarrow 0} I_d / \sqrt[k]{d}$.

Certain geometric terminology concerning conic sections has been taken over, via the Dupin indicatrix, to surfaces. Given a conic section in the plane, the directions which we have chosen as the x - and y - axes are called its **principal axes**. Consequently, the unit vectors $X_1, X_2 \in M_p$ (that is, the unit eigenvectors for $-dv: M_p \rightarrow M_p$) are called the **principal vectors**. They are really defined



only up to sign, so it is often more convenient to speak of the **principal directions**; moreover, if $k_1 = k_2$, then all unit vectors are to be considered to be principal. The eigenvalues k_1 and k_2 are called the **principal curvatures** at p . We have already met these vectors and curvatures in Volume II, and we recall that if $X = (\cos \theta)X_1 + (\sin \theta)X_2$ is any other unit vector, then

$$\begin{aligned} \text{(II)} \quad \langle -dv(X), X \rangle &= \langle k_1(\cos \theta)X_1 + k_2(\sin \theta)X_2, (\cos \theta)X_1 + (\sin \theta)X_2 \rangle \\ &= k_1 \cos^2 \theta + k_2 \sin^2 \theta, \end{aligned}$$

which shows that k_1, k_2 are the minimum and maximum of $\langle -d\nu(X), X \rangle$ for unit vectors $X \in M_p$. Recall also that we defined

$$K(p) = \text{Gaussian curvature at } p = k_1 \cdot k_2$$

$$H(p) = \text{mean curvature at } p = \frac{1}{2}(k_1 + k_2).$$

A surface M is called **flat at** p if $K(p) = 0$. So p is a flat point if and only if p is either parabolic or planar.

For hyperbolas, there are two other important lines, the *asymptotes* (the dashed lines in the previous figure). The unit vectors which point along these lines in the Dupin indicatrix are called the **asymptotic directions**. If the hyperbola has the equation

$$k_1 x^2 + k_2 y^2 = \pm 1 \quad (k_1, k_2 \text{ of different signs}),$$

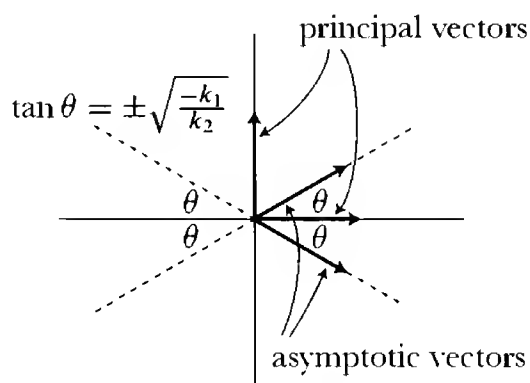
then the equation of the asymptotic lines $y = mx$ is found by noting that for large x the point (x, mx) is almost on the hyperbola, so

$$k_1 x^2 + k_2 m^2 x^2 \text{ is close to } \pm 1 \implies k_1 + k_2 m^2 \text{ is close to } 0,$$

and hence $m = \pm \sqrt{-k_1/k_2}$. On the other hand, if we consider a unit vector $X = (\cos \theta)X_1 + (\sin \theta)X_2$, then formula (II) gives

$$\langle -d\nu(X), X \rangle = k_1 \cos^2 \theta + k_2 \sin^2 \theta,$$

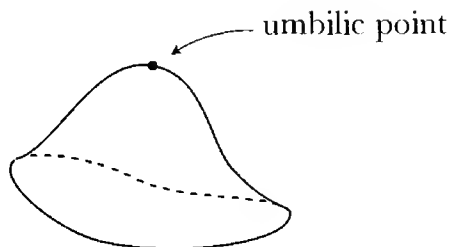
and this clearly equals 0 precisely when $\tan \theta = \pm \sqrt{-k_1/k_2}$. We therefore see that the vector $X \in M_p$ points along an asymptotic direction if and only if $\langle d\nu(X), X \rangle = 0$, and hence $\text{II}(X, X) = 0$.



Asymptotic directions do not exist at elliptic points, while there are two distinct asymptotic directions at hyperbolic points, and these directions are bisected by the principal directions. At a parabolic point there is only *one* asymptotic

direction—the principal direction with principal curvature 0. At planar points all directions are both principal and asymptotic. It is also clear that the asymptotic directions are perpendicular precisely when $k_1 = -k_2$, or $H = 0$.

Finally, there is one more important term, which describes a point where the Dupin indicatrix is actually a circle, so that the principal curvatures are equal, and all directions are principal directions. A point where all directions are principal is called an **umbilic** or **navel point**. This rather gross anatomical metaphor



is meant to suggest that the surface is very round at the point, like a sphere, on which all points are umbilics. Notice, however, that our definition also makes planar points umbilics, which turns out to be a convenient arrangement. At an umbilic, the map $-dv$ is just multiplication by some number k ; equation (I) therefore shows that

$$l_{ij} = k g_{ij} \quad \text{at an umbilic.}$$

At this stage it seems reasonable to begin asking to what extent the functions l, m, n describe f globally. The simplest question we can ask concerns surfaces f all of whose points are planar. Just as a curve with everywhere 0 curvature is a straight line, so we would expect a surface with $l = m = n = 0$ everywhere to be a plane. This is easy to prove. For, $l_{ij} = 0$ means $\langle -N_i, f_j \rangle = 0$; since N_i is a linear combination of f_1, f_2 , this means that $-N_i = 0$, so N is constant. Therefore $\langle f, N \rangle_i = \langle f_i, N \rangle + \langle f, N_i \rangle = 0 + 0$, and hence $\langle f, N \rangle = b$, where N is a constant vector and b is a constant number. This is just the equation of a plane.

It would next seem reasonable to prove an analogue of the fact that a circle is the only plane curve with constant κ . Here the situation is a little different, however: we cannot expect to characterize a sphere totally in terms of l, m, n , because we have not picked out a preferred parameterization; the functions E, F, G must also play a role. In fact, the simplest criterion to consider is that all points of f be umbilics. This means that

$$l = k E, \quad m = k F, \quad n = k G,$$

for a certain function k . Of course, in the case of a sphere, k is constant, but we do not even have to assume that. Although the following analysis is quite easy, it is worth recording as a theorem, which also includes the result about surfaces with all points planar.

2. THEOREM. If $M \subset \mathbb{R}^3$ is a connected surface such that every point is an umbilic, then M is part of a plane or a sphere.

PROOF. Choose an immersion $f: U \rightarrow M$. By assumption, we have $\langle -N_i, f_j \rangle = \langle k f_i, f_j \rangle$. Since N_i is a linear combination of f_1, f_2 , we thus have

$$(1) \quad N_i = -k f_i.$$

Consequently,

$$N_{ij} = -k_j f_i - k f_{ij}.$$

Since $N_{ij} = N_{ji}$, we obtain

$$-k_j f_i - k f_{ij} = -k_i f_j - k f_{ij},$$

and hence $k_i f_j = k_j f_i$. Setting $i = 1, j = 2$, and using linear independence of f_1, f_2 , we obtain $k_i = 0$, so k is constant. Thus equation (1) gives

$$(2) \quad N = -k f + v_0 \quad \text{for some } v_0 \in \mathbb{R}^3.$$

If $k = 0$, then as we already showed above, f lies in a plane. If $k \neq 0$, we have

$$f - \frac{v_0}{k} = \frac{-N}{k} \implies \left| f - \frac{v_0}{k} \right|^2 = \frac{1}{k^2},$$

so f lies in a sphere of radius $1/|k|$. Simple supplementary considerations then allow one to deduce the stated result. ♦

We now want to carry the analogy with curves still further, and see whether every immersion $f: U \rightarrow \mathbb{R}^3$ is described completely by the corresponding g_{ij} and l_{ij} . Of course, we only expect g_{ij} and l_{ij} to determine f up to proper Euclidean motions (translations followed by rotations [elements of $\text{SO}(3)$]), since g_{ij} and l_{ij} are already “invariant under proper Euclidean motions”—if A is a proper Euclidean motion, then the g_{ij} and l_{ij} for $A \circ f$ are the same as those for f . In the theory of curves we showed that κ and τ formed a complete set of invariants for a curve up to translations and rotations, by showing that they were a complete set of invariants up to rotation for the function

$s \mapsto (\mathbf{t}(s), \mathbf{n}(s), \mathbf{b}(s))$; this was accomplished by using the Serret-Frenet formulas, which are differential equations for $(\mathbf{t}, \mathbf{n}, \mathbf{b})$, involving only κ and τ . In the case of surfaces, we have the three vectors (f_1, f_2, N) , and so we want first to express the derivatives of each of these vectors as linear combinations of these same three vectors.

We begin by considering the f_{ik} , which we want to write as

$$(1) \quad f_{ik} = \sum_{h=1}^2 A_{ik}^h f_h + B_{ik} N.$$

First we will worry about finding the A_{ik}^h , which amounts to finding the $\langle f_{ik}, f_j \rangle$. To do this, we should obviously begin with the definition $\langle f_i, f_j \rangle = g_{ij}$ and differentiate, to get

$$\langle f_{ik}, f_j \rangle + \langle f_i, f_{jk} \rangle = g_{ij,k} \quad [\text{here } g_{ij,1} = D_1 g_{ij} = \frac{\partial g_{ij}}{\partial s}, \text{ etc.}].$$

Now comes the familiar old switcheroo: We also have

$$\begin{aligned} \langle f_{ji}, f_k \rangle + \langle f_j, f_{ki} \rangle &= g_{jk,i} \\ \langle f_{kj}, f_i \rangle + \langle f_k, f_{ij} \rangle &= g_{ki,j}; \end{aligned}$$

adding the first two of these three equations and subtracting the third, we get

$$\begin{aligned} \langle f_{ik}, f_j \rangle &= \frac{1}{2}(g_{ij,k} + g_{jk,i} - g_{ik,j}) \\ &= [ik, j], \end{aligned}$$

where $[ik, j]$ is the Christoffel symbol for the metric $I_f = f^*\langle \ , \ \rangle$ on U with respect to the standard coordinate system (s, t) on \mathbb{R}^2 . Plugging back into (1) we have

$$[ik, j] = \langle f_{ik}, f_j \rangle = \sum_{h=1}^2 A_{ik}^h g_{hj},$$

and, of course, we can solve explicitly for A_{ik}^h , using the g^{ij} :

$$A_{ik}^\rho = \sum_{j=1}^2 g^{\rho j} [ik, j] = \Gamma_{ik}^\rho.$$

There is no problem finding the B_{ik} , since we have already introduced a name for them:

$$B_{ik} = \langle f_{ik}, N \rangle = -\langle N_t, f_k \rangle = l_{ik}.$$

We have thus found that

$$(*) \quad f_{ik} = \sum_{h=1}^2 \Gamma_{ik}^h f_h + l_{ik} N \quad \text{The Gauss Formulas.}$$

The reader can easily check that these equations are indeed precisely the Gauss Formulas on page 4. Of course, in our present derivation, it is unnecessary to use the ∇ operator on the image of f , or even to have it defined; the only ∇ operator one needs to know about is the one for \mathbb{R}^3 , and the Γ 's just appear as weird combinations of the g_{ij} and their derivatives. (It is hard to see why these formulas should be named after Gauss, for he never explicitly solves for the f_{ik} . The closest results he writes down are certain formulas [for m, m', m'', n, n', n'' , on pg. II.91] equivalent to the equations $\langle f_{ik}, f_j \rangle = [ik, j]$.)

We next want to express N_i in terms of f_1, f_2, N , as

$$N_i = \sum_{h=1}^2 C_i^h f_h + 0 \cdot N.$$

Once again, there is no problem here, since we have already introduced a name for the relevant inner products. We have

$$l_{ij} = \langle -N_i, f_j \rangle = - \sum_{h=1}^2 C_i^h g_{hj},$$

and consequently

$$C_i^\rho = - \sum_{j=1}^2 g^{\rho j} l_{ij}.$$

Introducing new symbols l_i^h , we can therefore write

$$(**) \quad N_i = - \sum_{h=1}^2 \left(\sum_{j=1}^2 g^{hj} l_{ij} \right) f_h = - \sum_{h=1}^2 l_i^h f_h \quad \text{The Weingarten Equations.}$$

Of course, equations (**) amount to little more than the definition of l_{ij} and l_i^h , but in the classical literature it is always precisely these equations which are called the Weingarten equations.

The Gauss and Weingarten equations constitute an exact analogue of the Serret-Frenet formulas for a curve—the derivatives of f_1, f_2, N have been

expressed in terms of these same vectors, and only the g_{ij} and l_{ij} enter. We thus seem to be in a good position to produce an immersion f with given g_{ij} and l_{ij} : we should first solve the Gauss and Weingarten equations for f_1, f_2, N , and then solve for f . However, these equations are *partial* differential equations (just 15 in all, for the 9 component functions f_i^j, n^j), and we know that these equations have solutions only if certain compatibility conditions are satisfied. These conditions are given explicitly on pg. I.187, but there is no need to turn back to them: the required conditions are obtained simply by setting mixed partial derivatives equal, and substituting the original equation into the results so obtained. We will now derive these conditions explicitly.

We begin by using (*) to compute

$$\begin{aligned}
 f_{ikj} &= \sum_{h=1}^2 \Gamma_{ik,j}^h f_h + \sum_{h=1}^2 \Gamma_{ik}^h f_{hj} + l_{ik,j} N + l_{ik} N_j \\
 &\quad [\text{here } \Gamma_{ik,1}^h = \partial \Gamma_{ik}^h / \partial s, \text{ etc.}] \\
 &= \sum_{\rho=1}^2 \Gamma_{ik,j}^\rho f_\rho + \sum_{h=1}^2 \Gamma_{ik}^h \left(\sum_{\rho=1}^2 \Gamma_{hj}^\rho f_\rho + l_{hj} N \right) \\
 &\quad + l_{ik,j} N - l_{ik} \left(\sum_{\rho=1}^2 l_j^\rho f_\rho \right), \quad \text{using (*) and (**).}
 \end{aligned}$$

Setting $f_{ikj} = f_{ijk}$, and using linear independence of f_1, f_2, N , we have

$$(A) \quad \Gamma_{ik,j}^\rho - \Gamma_{ij,k}^\rho + \sum_{h=1}^2 (\Gamma_{ik}^h \Gamma_{hj}^\rho - \Gamma_{ij}^h \Gamma_{hk}^\rho) = l_{ik} l_j^\rho - l_{ij} l_k^\rho$$

$$(B) \quad l_{ik,j} - l_{ij,k} + \sum_{h=1}^2 \Gamma_{ik}^h l_{hj} - \sum_{h=1}^2 \Gamma_{ij}^h l_{hk} = 0.$$

In this mess, some things should be looking familiar. Indeed, comparing with pg. II.188, we see that equation (A) says that

$$R^\rho_{kji} = l_{ik} l_j^\rho - l_{ij} l_k^\rho.$$

which is equivalent to

$$\begin{aligned}
 (A') \quad R_{hkji} &= \sum_{\rho=1}^2 g_{h\rho} R^{\rho}_{kji} = \sum_{\rho=1}^2 g_{h\rho} (l_j^{\rho} l_{ik} - l_k^{\rho} l_{ij}) \\
 &= \sum_{\rho=1}^2 g_{h\rho} \left(\sum_{\sigma=1}^2 g^{\rho\sigma} l_{\sigma j} l_{ik} - \sum_{\sigma=1}^2 g^{\rho\sigma} l_{\sigma k} l_{ij} \right) \\
 &\quad \text{by the definition of } l_j^{\rho} \text{ in } (**) \\
 &= l_{hj} l_{ik} - l_{hk} l_{ij}.
 \end{aligned}$$

A special case is

$$R_{1212} = l_{11} l_{22} - l_{12} l_{12} = ln - m^2 \quad \begin{array}{l} \text{Gauss' Equation} \\ \text{(Gauss' Theorema Egregium).} \end{array}$$

This really is equivalent to Gauss' Theorema Egregium, for it says (cf. pg. II.190) that

$$\left\langle R \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial t} \right) \frac{\partial f}{\partial t}, \frac{\partial f}{\partial s} \right\rangle = ln - m^2,$$

and hence that the intrinsically defined Gaussian curvature K is given by

$$K = \frac{\left\langle R \left(\frac{\partial f}{\partial s}, \frac{\partial f}{\partial t} \right) \frac{\partial f}{\partial t}, \frac{\partial f}{\partial s} \right\rangle}{\left\langle \frac{\partial f}{\partial s}, \frac{\partial f}{\partial s} \right\rangle \left\langle \frac{\partial f}{\partial t}, \frac{\partial f}{\partial t} \right\rangle - \left\langle \frac{\partial f}{\partial s}, \frac{\partial f}{\partial t} \right\rangle^2} = \frac{ln - m^2}{EG - F^2},$$

the final expression being the Gaussian curvature as originally (extrinsically) defined for a surface in \mathbb{R}^3 . In Volume II we gave a simpler looking proof of this result, but the present proof is philosophically more satisfying, since it relies only on standard techniques for dealing with a system of partial differential equations. Notice that our proof of Theorem I-6 was basically the same, since it used the fact that $R(X, Y)Z$ measures the difference of $\nabla_X \nabla_Y Z$ and $\nabla_Y \nabla_X Z$.

It is easy to see that all other cases of (A') are equivalent to this particular one, or are trivial, because of the identities

$$R_{ijkl} = -R_{jikl} \quad \text{and} \quad R_{klij} = R_{ijkl},$$

which always hold (pp. II.194ff.), and the fact that the right side of (A') has the same symmetry properties.

Now let us take a look at (B). If $j = k$ it says nothing. Moreover, the equation for $j = 2, k = 1$ is equivalent to the one for $j = 1, k = 2$. So we take the latter pair for j and k , and let $i = 1$ or 2 , obtaining

$$(B') \quad \begin{cases} l_{12,1} - l_{11,2} + \sum_{h=1}^2 \Gamma_{12}^h l_{h1} - \sum_{h=1}^2 \Gamma_{11}^h l_{h2} = 0 \\ l_{22,1} - l_{21,2} + \sum_{h=1}^2 \Gamma_{22}^h l_{h1} - \sum_{h=1}^2 \Gamma_{21}^h l_{h2} = 0 \end{cases} \quad \begin{array}{l} \text{The Codazzi-Mainardi} \\ \text{Equations.} \end{array}$$

It is easy to see (Problem 1) that these equations can be derived from the ones given in Corollary 1-12. [Note also that equations (A') and (B') are precisely what the classical tensor analysis equations (11) and (12) on page 16 become in this case.]

There is still one more set of equations which we must consider, obtained by setting $N_{ij} = N_{ji}$. However (Problem 2), it turns out that these reduce to the Codazzi-Mainardi equations. We have thus found altogether three conditions which must be satisfied, and our general theory (Theorem I.6-1) tells us that these are the only conditions we need. We are all ready for a theorem.

3. FUNDAMENTAL THEOREM OF SURFACE THEORY (BONNET; 1867). Let $U \subset \mathbb{R}^2$ be a convex open set containing $(0, 0)$.

(1) Let $f, \bar{f}: U \rightarrow \mathbb{R}^3$ be two immersions, and define

$$\begin{aligned} g_{ij} &= \langle f_i, f_j \rangle & \bar{g}_{ij} &= \langle \bar{f}_i, \bar{f}_j \rangle \\ N &= \frac{f_1 \times f_2}{\sqrt{g_{11}g_{22} - g_{12}^2}} & \bar{N} &= \frac{\bar{f}_1 \times \bar{f}_2}{\sqrt{\bar{g}_{11}\bar{g}_{22} - \bar{g}_{12}^2}} \\ l_{ij} &= \langle -N_i, f_j \rangle = \langle N, f_{ij} \rangle & \bar{l}_{ij} &= \langle -\bar{N}_i, \bar{f}_j \rangle = \langle \bar{N}, \bar{f}_{ij} \rangle. \end{aligned}$$

Suppose that $g_{ij} = \bar{g}_{ij}$ and $l_{ij} = \bar{l}_{ij}$ on U . Then there is a proper Euclidean motion A such that $\bar{f} = A \circ f$.

(2) Let g_{ij} and l_{ij} ($i, j = 1, 2$) be functions on U which satisfy

- (i) $g_{ij} = g_{ji}$ and $l_{ij} = l_{ji}$, and (g_{ij}) is positive definite on U , so that we can define corresponding g^{ij} and Γ_{ij}^k

(ii) Gauss' Equation:

$$\begin{aligned} l_{11}l_{22} - (l_{12})^2 &= R_{1212} \\ &= \sum_{\rho=1}^2 g_{1\rho} \left(\Gamma_{22,1}^\rho - \Gamma_{21,2}^\rho + \sum_{h=1}^2 (\Gamma_{22}^h \Gamma_{h1}^\rho - \Gamma_{21}^h \Gamma_{h2}^\rho) \right) \end{aligned}$$

(iii) The Codazzi-Mainardi Equations:

$$\begin{aligned} l_{12,1} - l_{11,2} + \sum_{h=1}^2 \Gamma_{12}^h l_{h1} - \sum_{h=1}^2 \Gamma_{11}^h l_{h2} &= 0 \\ l_{22,1} - l_{21,2} + \sum_{h=1}^2 \Gamma_{22}^h l_{h1} - \sum_{h=1}^2 \Gamma_{21}^h l_{h2} &= 0. \end{aligned}$$

Then there is an immersion $f: U \rightarrow \mathbb{R}^3$ such that

$$\begin{aligned} g_{ij} &= \langle f_i, f_j \rangle \\ l_{ij} &= \langle -N_i, f_j \rangle = \langle N, f_{ji} \rangle, \quad \text{for } N = \frac{f_1 \times f_2}{\sqrt{g_{11}g_{22} - g_{12}^2}}. \end{aligned}$$

PROOF. Let us adopt the more systematic notation

$$\begin{aligned} \mathbf{v}_1 &= f_1, \quad \mathbf{v}_2 = f_2, \quad \mathbf{v}_3 = N \\ \bar{\mathbf{v}}_1 &= \bar{f}_1, \quad \bar{\mathbf{v}}_2 = \bar{f}_2, \quad \bar{\mathbf{v}}_3 = \bar{N}. \end{aligned}$$

To prove (I), we first choose a rotation $B \in \text{SO}(3)$ such that

$$B(\mathbf{v}_\alpha(0,0)) = \bar{\mathbf{v}}_\alpha(0,0) \quad \alpha = 1, 2, 3.$$

This is possible because $g_{ij}(0) = \bar{g}_{ij}(0)$ for $i, j = 1, 2$, and because the two triples of vectors $(\mathbf{v}_1(0,0), \mathbf{v}_2(0,0), \mathbf{v}_3(0,0))$ and $(\bar{\mathbf{v}}_1(0,0), \bar{\mathbf{v}}_2(0,0), \bar{\mathbf{v}}_3(0,0))$ are both positively oriented, with the third vector perpendicular to the first two. If we let $\tilde{f} = B \circ f$, then it is easy to see that

$$\begin{aligned} \tilde{g}_{ij} &= g_{ij} = \bar{g}_{ij} \\ \tilde{\mathbf{v}}_3 &= B \circ \mathbf{v}_3 \\ \tilde{l}_{ij} &= l_{ij} = \bar{l}_{ij}. \end{aligned}$$

We claim that the maps

$$(\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \bar{\mathbf{v}}_3), (\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \tilde{\mathbf{v}}_3): U \rightarrow \mathbb{R}^3,$$

which we know are equal at $(0, 0)$, are actually equal everywhere. To prove this, we recall that the Gauss formulas and the Weingarten equations give

$$(***) \quad \begin{cases} \bar{\mathbf{v}}_{i,k}(s, t) = \sum_{h=1}^2 \bar{\Gamma}_{ik}^h(s, t) \bar{\mathbf{v}}_h(s, t) + \bar{l}_{ik}(s, t) \bar{\mathbf{v}}_3(s, t) & i = 1, 2 \\ \bar{\mathbf{v}}_{3,k}(s, t) = - \sum_{h=1}^2 \left(\sum_{j=1}^2 \bar{g}^{hj}(s, t) \bar{l}_{kj}(s, t) \right) \bar{\mathbf{v}}_h(s, t) \end{cases}$$

for the $\bar{\mathbf{v}}_\alpha$, while for the $\tilde{\mathbf{v}}_\alpha$ we obtain the corresponding equations with $\tilde{\Gamma}_{ik}^h$, \tilde{l}_{ik} and \tilde{g}^{hj} . But $\tilde{l}_{ik} = \bar{l}_{ik}$, and since $\tilde{g}_{ij} = \bar{g}_{ij}$ we also have $\tilde{g}^{hj} = \bar{g}^{hj}$ and $\tilde{\Gamma}_{ik}^h = \bar{\Gamma}_{ik}^h$. So the two maps $(\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \bar{\mathbf{v}}_3)$ and $(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \tilde{\mathbf{v}}_3)$ satisfy the *same* equations (***) and have the same values at $(0, 0)$. Therefore they must be equal on U (Theorem I.6-1). But this means that \tilde{f} and $\bar{f} = B \circ f$ have the same partial derivatives, and therefore differ by a constant vector. Consequently, there is a translation $T: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with $\tilde{f} = T \circ \bar{f} = (T \circ B) \circ f$.

To prove (2), we use Theorem I.6-1 to conclude that equation (***), written in terms of the given g_{ij} and l_{ij} , has a solution $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3: U \rightarrow \mathbb{R}^3$ with any desired initial conditions; we have already seen that the required conditions in Theorem I.6-1 amount precisely to Gauss' equation and the Codazzi-Mainardi equations. Moreover, the functions \mathbf{v}_α can be defined on all of U because the equations (***) are linear (compare pg. I.165). Since (g_{ij}) is positive definite at $(0, 0)$, there is a solution for which the following conditions are satisfied at $(s, t) = (0, 0)$:

- (a) $\langle \mathbf{v}_i(s, t), \mathbf{v}_j(s, t) \rangle = g_{ij}(s, t) \quad i, j = 1, 2$
- (b) $\langle \mathbf{v}_i(s, t), \mathbf{v}_3(s, t) \rangle = 0 \quad i = 1, 2$
- (c) $|\mathbf{v}_3(s, t)| = 1$
- (d) $(\mathbf{v}_1(s, t), \mathbf{v}_2(s, t), \mathbf{v}_3(s, t))$ is positively oriented.

We will show that conditions (a)–(d) actually hold at all points of U .

Our equation (***) for $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ gives the equations

$$\begin{aligned} \text{(A)} \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle_k &= \langle \mathbf{v}_{i,k}, \mathbf{v}_j \rangle + \langle \mathbf{v}_i, \mathbf{v}_{j,k} \rangle \\ &= \sum_{h=1}^2 \Gamma_{ik}^h \langle \mathbf{v}_h, \mathbf{v}_j \rangle + \sum_{h=1}^2 \Gamma_{jk}^h \langle \mathbf{v}_h, \mathbf{v}_i \rangle + l_{ik} \langle \mathbf{v}_3, \mathbf{v}_i \rangle + l_{jk} \langle \mathbf{v}_3, \mathbf{v}_j \rangle \end{aligned}$$

for $i, j = 1, 2$, as well as

$$\begin{aligned}
 \text{(B)} \quad \langle \mathbf{v}_i, \mathbf{v}_3 \rangle_k &= \langle \mathbf{v}_{i,k}, \mathbf{v}_3 \rangle + \langle \mathbf{v}_i, \mathbf{v}_{3,k} \rangle \\
 &= l_{ik} - \sum_{h=1}^2 \left(\sum_{j=1}^2 g^{hj} l_{kj} \right) \langle \mathbf{v}_i, \mathbf{v}_h \rangle
 \end{aligned}$$

and

$$\text{(C)} \quad \langle \mathbf{v}_3, \mathbf{v}_3 \rangle_k = 2 \langle \mathbf{v}_{3,k}, \mathbf{v}_3 \rangle = 0.$$

[Equations (A)–(C) all hold for $k = 1, 2$.]

But we also have

$$\begin{aligned}
 g_{ij,k} &= [ik, j] + [jk, i] \quad (\text{by pg. I.331}) \\
 &= \sum_{h=1}^2 \Gamma_{ik}^h g_{hj} + \sum_{h=1}^2 \Gamma_{jk}^h g_{hi}.
 \end{aligned}$$

This shows that the set of equations (A)–(C) are satisfied both by

the set of functions: $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ ($j = 1, 2$), $\langle \mathbf{v}_3, \mathbf{v}_1 \rangle$, $\langle \mathbf{v}_3, \mathbf{v}_2 \rangle$, $\langle \mathbf{v}_3, \mathbf{v}_3 \rangle$

and by

the set of functions: g_{ij} ($j = 1, 2$), 0 , 0 , 1 .

Moreover, we chose the \mathbf{v}_i so that these two collections of functions have the same value at $(0, 0)$. It follows that they have the same values on all of U . In other words, equations (a)–(c) hold on all of U . Moreover, (a) and (b) [and non-singularity of (g_{ij})] imply that $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ are always linearly independent. So condition (d) at $(0, 0)$ implies condition (d) everywhere.

We now claim that there is a function $f: U \rightarrow \mathbb{R}^3$ satisfying $f_i = \mathbf{v}_i$. In order to prove this, we just have to show that $\mathbf{v}_{i,j} = \mathbf{v}_{j,i}$. But this follows from (**), by symmetry of the Γ_{ik}^h and l_{ik} . We now have $\langle f_i, f_j \rangle = g_{ij}$ by (a). Moreover, (b)–(d) then show that $\mathbf{v}_3 = n$. Consequently,

$$\langle f_{ij}, n \rangle = \langle \mathbf{v}_{i,j}, \mathbf{v}_3 \rangle = l_{ij}$$

by (**), together with (b) and (c). ♦

Theorem 3 is exactly the sort of result we would want if we were primarily interested in immersions $f: U \rightarrow \mathbb{R}^3$. But what we really want to study are submanifolds of \mathbb{R}^3 , without relying on a particular choice of a parameterization. For example, let us consider two surfaces $M, \bar{M} \subset \mathbb{R}^3$ and a diffeomorphism $\phi: M \rightarrow \bar{M}$. We would like to have conditions which insure that ϕ is the restriction to M of some proper Euclidean motion. If we arbitrarily choose some immersion $f: U \rightarrow M \subset \mathbb{R}^3$, and let $\bar{f}: U \rightarrow \bar{M} \subset \mathbb{R}^3$ be $\bar{f} = \phi \circ f$, then Theorem 2 tells us that such a proper Euclidean motion exists if the g_{ij} and l_{ij} for f equal the \bar{g}_{ij} and \bar{l}_{ij} for \bar{f} . Now the individual functions g_{ij} and l_{ij} for f do not have an “invariant meaning”: given a submanifold $M \subset \mathbb{R}^3$, we cannot, for example, find functions γ_{ij} on M so that every $f: U \rightarrow M$ has its g_{ij} ’s given simply by $g_{ij} = \gamma_{ij} \circ f$. Fortunately, however, the tensors

$$\begin{aligned} g_{11} ds \otimes ds + g_{12} ds \otimes dt + g_{21} dt \otimes ds + g_{22} dt \otimes dt \\ l_{11} ds \otimes ds + l_{12} ds \otimes dt + l_{21} dt \otimes ds + l_{22} dt \otimes dt \end{aligned}$$

do have an invariant meaning: they are just f^*I and f^*II . So we can formulate the first part of Theorem 3 for submanifolds:

4. COROLLARY. Let $M, \bar{M} \subset \mathbb{R}^3$ be two connected oriented surfaces imbedded in \mathbb{R}^3 , let $\nu: M \rightarrow S^2 \subset \mathbb{R}^3$ and $\bar{\nu}: \bar{M} \rightarrow S^2 \subset \mathbb{R}^3$ be the unit normal vector fields determined by the orientations, and let I, II and \bar{I}, \bar{II} be the first and second fundamental forms for M and \bar{M} (the forms \bar{I} and \bar{II} being defined with respect to ν and $\bar{\nu}$, respectively). Let $\phi: M \rightarrow \bar{M}$ be an orientation preserving diffeomorphism which preserves the first and second fundamental forms,

$$\begin{aligned} \phi^*\bar{I} &= I \quad (\text{i.e., } \phi \text{ is an isometry}) \\ \phi^*\bar{II} &= II. \end{aligned}$$

Then there is a proper Euclidean motion A such that $\phi = A|_M$ and $A_*\nu = \bar{\nu}$.

PROOF. Let $f: U \rightarrow M \subset \mathbb{R}^3$ be an orientation preserving immersion, and let $\bar{f} = \phi \circ f: U \rightarrow \bar{M} \subset \mathbb{R}^3$; the immersion \bar{f} is also orientation preserving, since ϕ is. The \bar{g}_{ij} for \bar{f} are the coefficients, with respect to the standard coordinate system (s, t) , of

$$\bar{f}^*\bar{I} = (\phi \circ f)^*\bar{I} = f^*\phi^*\bar{I} = f^*I.$$

Consequently, $\bar{g}_{ij} = g_{ij}$. Similarly, since \bar{f} is orientation preserving, the \bar{l}_{ij} are the coefficients of $\bar{f}^*\bar{II} = f^*II$; since f is also orientation preserving, we find that $\bar{l}_{ij} = l_{ij}$. By Theorem 3, there is some proper Euclidean motion A such that $\phi = A$ on $f(U)$. If we choose immersions $\{f_\alpha: U_\alpha \rightarrow M\}$ whose images

cover M , it is easy to see that the corresponding A_α must all be the same proper Euclidean motion A . ♦

We also want to formulate the existence part of Theorem 3 for manifolds, rather than immersions. So we consider an oriented surface M with a Riemannian metric $\langle \cdot, \cdot \rangle$ [corresponding to the g_{ij}] and a symmetric tensor S covariant of order 2 [corresponding to the l_{ij}]. In the previous chapter we have already seen how to give an invariant version of Gauss' equation and the Codazzi-Mainardi equations. This allows us to state

5. COROLLARY. Let $(M, \langle \cdot, \cdot \rangle)$ be an oriented Riemannian 2-manifold, with covariant derivative ∇ and curvature tensor R , and let S be a symmetric tensor on M , covariant of order 2. Suppose that S satisfies

(1) Gauss' Equation:

$$\langle R(X, Y)Y, X \rangle = S(X, X)S(Y, Y) - [S(X, Y)]^2$$

(2) The Codazzi-Mainardi Equations:

$$(\nabla_X S)(Y, Z) = (\nabla_Y S)(X, Z).$$

Then for any $p \in M$ there is a neighborhood U of p and an immersion $f: U \rightarrow \mathbb{R}^3$ such that

$$\begin{aligned} \langle \cdot, \cdot \rangle &= f^* \langle \cdot, \cdot \rangle \\ S &= f^* \text{II}, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the usual Riemannian metric on \mathbb{R}^3 and II is the second fundamental form on $f(U)$ defined in terms of the unit normal field ν which is determined by the orientation that $f(U)$ gets from the orientation on $U \subset M$.

PROOF. Left to the reader. ♦

Unlike Corollary 4, where a global result comes almost automatically, in Corollary 5 we cannot generally choose U to be all of M . As an example, we take the torus $S^1 \times S^1$ with a flat metric $\langle \cdot, \cdot \rangle$ [pg. II.179] and let $S = 0$. The Gauss and Codazzi-Mainardi equations are trivially satisfied. But the only connected submanifolds of \mathbb{R}^3 with $\text{II} = 0$ everywhere are subsets of a plane (Theorem 2), so we certainly cannot find an immersion $f: S^1 \times S^1 \rightarrow \mathbb{R}^3$ with $f^* \text{II} = S = 0$ everywhere. On the other hand (Problem 3), we *can* take U to be all of M in Corollary 5 when M is simply-connected.

Now that we have adequately documented the importance of the second fundamental form in surface theory, we will take this opportunity to slip in something new. The reader has perhaps already surmised with subconscious dread that there is a third fundamental form, and even yet higher numbered monsters, but these hogey men turn out to be very nicely behaved creatures which are in no way to be feared. For a submanifold $M \subset \mathbb{R}^3$, with unit normal $v: M \rightarrow S^2 \subset \mathbb{R}^3$, we define the **third fundamental form** III of M by

$$\begin{aligned} \text{III}(p)(v_p, w_p) &= \langle -dv(v_p), -dv(w_p) \rangle \\ &= \langle dv(v_p), dv(w_p) \rangle \quad v_p, w_p \in M_p. \end{aligned}$$

Similarly, if $f: U \rightarrow \mathbb{R}^3$ is an immersion (for $U \subset \mathbb{R}^2$ open), we define III_f by

$$\begin{aligned} \text{III}_f(s, t)(v, w) &= \langle dN_f(v), dN_f(w) \rangle \\ &= \langle (N_f)_*(v), (N_f)_*(w) \rangle \quad v, w \in \mathbb{R}^2_{(s, t)}. \end{aligned}$$

This is equivalent to defining $\text{III}_f = f^*\text{III}$, where III is the third fundamental form for image f . Remembering that $dv: M_p \rightarrow M_p$ is self-adjoint (Theorem 1-8 or Theorem II.3-1), we see that

$$\text{III}(p)(v_p, w_p) = \langle (dv)^2(v_p), w_p \rangle.$$

This suggests defining

$$\text{IV}(p)(v_p, w_p) = \langle (dv)^3(v_p), w_p \rangle,$$

and so forth. There is no notational way to write down the general definition, since no one has ever addressed the burning question of how we should indicate the n^{th} Roman numeral (come to think of it, no one even knows how to write down arbitrarily large Roman numerals). But that doesn't matter very much, especially as all these forms are expressible in terms of I and II anyway:

6. PROPOSITION. For a surface $M \subset \mathbb{R}^3$ we have

$$\text{III} - 2H \cdot \text{II} + K \cdot \text{I} = 0.$$

(Similarly, for an immersion $f: U \rightarrow \mathbb{R}^3$, we have

$$\text{III}_f - 2H \cdot \text{II}_f + K \cdot \text{I}_f = 0.$$

where $H(s, t)$ is the mean curvature of image f at $f(s, t)$, etc.)

PROOF. Remember the Cayley-Hamilton Theorem! The map $-dv: M_p \rightarrow M_p$ satisfies its characteristic polynomial $\chi(\lambda)$, which is given by

$$\chi(\lambda) = \lambda^2 - [\text{trace}(-dv)]\lambda + \det(-dv) = \lambda^2 - 2H\lambda + K.$$

Consequently,

$$(-dv)^2 - 2H(-dv) + K \cdot \text{identity} = 0 \quad \text{on } M_p.$$

Applying this equation to v_p , and taking the inner product with w_p , we obtain the desired result. ♦

It is clear that we can also express IV in terms of III and II, etc. Proposition 6 does not necessarily mean that III is not worth considering, for it is still a useful tool for expressing certain quantities. Suppose, for example, that we have an immersion $f: U \rightarrow \mathbb{R}^3$, with normal map $N_f (= v \circ f$ for the unit normal map v on image f). Since the normal map N_f plays such a vital role in describing the geometry of f , it is not at all unreasonable to ask what the first and second fundamental forms of N_f look like. Notice that in this instance we certainly want to explicitly consider the forms I_{N_f} and II_{N_f} for the map N_f : the image of N_f is just part of S^2 , so its first and second fundamental forms aren't very interesting.

7. PROPOSITION. Let $f: U \rightarrow \mathbb{R}^3$ be an immersion with normal map $N_f = v \circ f$. Then the third fundamental form of f is both the first fundamental form of N_f and the negative of the second fundamental form of N_f :

$$III_f = I_{N_f} = -II_{N_f}.$$

PROOF. Our original definition,

$$III_f(s, t)(v, w) = \langle (N_f)_*(v), (N_f)_*(w) \rangle,$$

shows that $III_f = I_{N_f}$. Since the unit normal vector at any point $p \in S^2$ is just p itself, it is also clear that the normal map of N_f is just N_f itself, so we have $N_{N_f} = N_f$. Thus

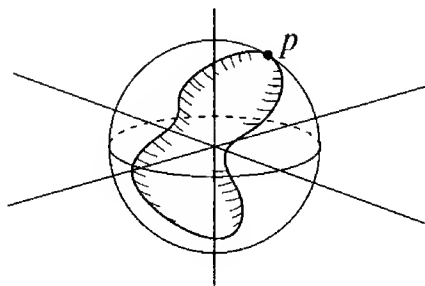
$$\begin{aligned} II_{N_f}(s, t)(v, w) &= \langle -(N_{N_f})_*(v), (N_f)_*(w) \rangle \\ &= \langle -(N_f)_*(v), (N_f)_*(w) \rangle. \quad \diamond \end{aligned}$$

This result will come in handy at one point in Chapter 9, but we have no more to say about III at present. Following the route set forth in the first chapter of Volume II, we will now briefly look at some global properties of surfaces which are related to positive curvature. At the very outset we note one respect in which the situation for surfaces in \mathbb{R}^3 is different from that of curves in \mathbb{R}^2 : Although

we are able to define a signed curvature κ for a curve c in \mathbb{R}^2 , the sign of κ depends on the “orientation” of c , and is reversed when we traverse c in the opposite direction; but for a surface $M \subset \mathbb{R}^3$, the Gaussian curvature K , which may also be positive or negative, does not depend on the orientation of M . We begin with a simple, but sometimes useful, observation.

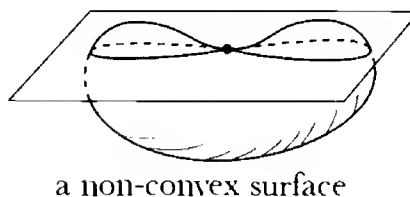
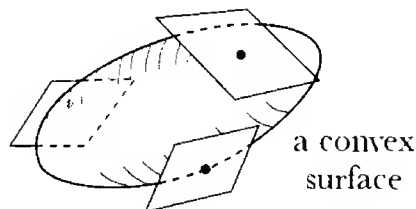
8. PROPOSITION. If M is a compact surface immersed in \mathbb{R}^3 , then there is at least one point $p \in M$ where $K(p) > 0$.

PROOF. The trick is to choose a point $p \in M$ whose distance from 0 is a maximum. Then M is even more curved at p than the sphere of radius $|p|$ —in fact, each principal curvature is $> 1/|p|$, by Proposition II.3-0, and the corresponding result for curves in \mathbb{R}^2 . Details are left as an exercise. ♦



This result gives us another way of seeing that the flat torus cannot be immersed in \mathbb{R}^3 (no matter what tensor S we choose on it).

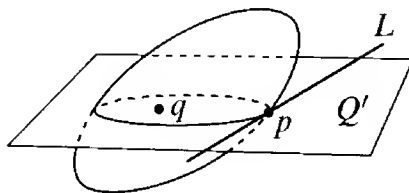
We now want to consider surfaces M with $K(p) > 0$ for *all* $p \in M$. We naturally hope to relate this condition to convexity, so a brief discussion of that concept is in order. We define an imbedded surface $M \subset \mathbb{R}^3$ to be **convex** if it lies on one side of each of its tangent planes. As in the case of curves, we would first like to relate this definition to the more common one.



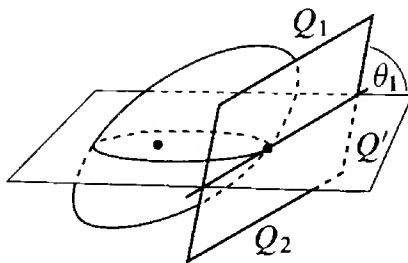
Any subset A of \mathbb{R}^3 is called **convex** if the line segment \overline{pq} from p to q is contained in A whenever $p, q \in A$. Suppose A is convex and p is a point in the boundary of A . A plane P containing p is called a **support plane** of A if A lies completely in one of the closed half-spaces into which P divides \mathbb{R}^3 .

9. PROPOSITION. If A is convex, and p is in the boundary of A , then there is at least one support plane P containing p .

PROOF. If A has no interior points, it lies in a plane, and the Theorem follows easily from the corresponding result, Proposition II.1-3, for subsets of \mathbb{R}^2 . If A has an interior point q , let Q be a plane containing q and p , and let L be a support line for $A \cap Q$ through p . This line L divides Q into two closed half-planes; let Q' be the one such that $Q' - L$ contains no points of $Q \cap A$.



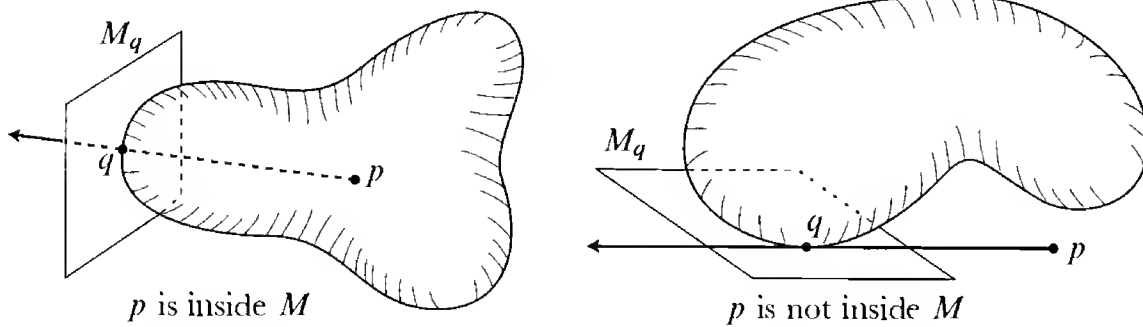
Now we will consider the various closed half-planes having L as their edge. Choose one side of Q' and consider angles θ such that the half-plane with L as its edge which makes an angle of θ with Q' on this side does not intersect A except along L (it may be that $\theta = 0$ is the only possibility). Let θ_1 be the least upper bound of all such θ , and let Q_1 be the half-plane with L as edge which makes an angle of θ_1 . Let Q_2 be the corresponding half-plane on the other side of Q' .



To prove the theorem, it clearly suffices to prove that the angle between Q_1 and Q_2 is $\geq \pi$. We note that there are points of A on planes arbitrarily close to Q_1 and Q_2 . If the angle between Q_1 and Q_2 were $< \pi$, then we could consider a suitable pair of such points, together with points of A in a whole neighborhood of q , and find that A must contain some point of L in its interior, which is impossible. ♦

We now want to show that a compact connected 2-dimensional submanifold M of \mathbb{R}^3 is convex if and only if the set A consisting of all points on M or inside M is a convex subset of \mathbb{R}^3 . Once again, we are assuming Corollary I.11-15, and the following easy consequence:

Suppose $M \subset \mathbb{R}^3$ is a compact connected surface, and l is a ray from p which intersects M at just one point $q \neq p$. Suppose, moreover, that l does *not* lie along the tangent plane of M at q . Then p is inside M .



10. PROPOSITION. Let $M \subset \mathbb{R}^3$ be a compact connected surface, and let A be the set of all points on or inside M . Then M is convex (that is, M lies on one side of each of its tangent planes) if and only if A is convex.

PROOF. Exactly like the proof of Proposition II.1-4. ♦

We are finally ready to relate convexity and curvature. If you have found yourself nodding drowsily at the rather obvious generalizations of old material which occupied the last few pages, it is time to wake up now, because our result for surfaces is *not* just an analogue of the result for curves.

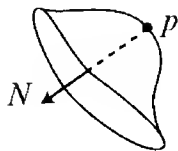
11. THEOREM (HADAMARD). (1) If M is a convex surface in \mathbb{R}^3 , then $K(p) \geq 0$ for all $p \in M$.

(2) Let M be a compact connected 2-manifold, and $f: M \rightarrow \mathbb{R}^3$ an immersion with $K(p) > 0$ for all $p \in M$. Then

- (i) The manifold M is orientable, and the normal map $N: M \rightarrow S^2 \subset \mathbb{R}^3$ is a diffeomorphism,
- (ii) The map $f: M \rightarrow \mathbb{R}^3$ is an imbedding, and $f(M)$ is convex.

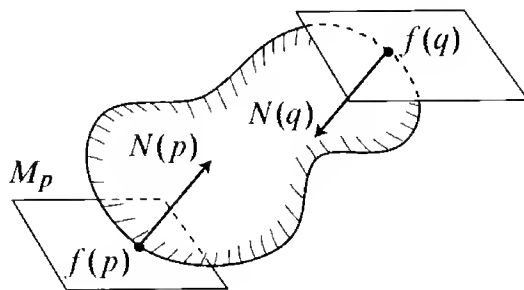
PROOF. The first part is immediate, for we have already seen that if $K(p) < 0$, then M lies on both sides of M_p .

To prove (2), we first recall that since $K(p) > 0$, points of M near p all lie on one side of M_p . We choose N so that it always points on this side. Since this

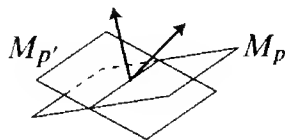


gives us a continuous choice of N , it also gives us an orientation of M . To show that $N: M \rightarrow S^2$ is a diffeomorphism, we note first that N_* is always one-one, since $K = \det N_*$. So $N(M) \subset S^2$ is open. It is also closed, since M is compact; so $N(M) = S^2$. Now we need to use a few properties of covering spaces. The fact that $N: M \rightarrow S^2$ is onto and locally one-one does not immediately imply that N is a covering space map; however it is an easy exercise (Problem 4) to show that this follows from the fact that M is compact. But S^2 is simply-connected, and therefore has no non-trivial covering spaces. So $N: M \rightarrow S^2$ is a diffeomorphism, and we have proved (i).

To prove (ii), we consider a point $p \in M$ and the tangent space $M_p \subset \mathbb{R}^3_{f(p)}$. At least some points of $f(M)$ lie on the same side of M_p as $N(p)$. Let $f(q)$ be



a point on this side which is furthest from M_p . Then clearly $N(q) = -N(p)$. Suppose that $f(p) = f(p')$ for some other $p' \in M$. Since $N: M \rightarrow \mathbb{R}^3$ is one-one, $N(p')$ cannot be either $N(p)$ or $-N(p) = N(q)$. So the tangent plane $M_{p'}$ must cross the tangent plane M_p . It is then easy to see that $f(M)$ must



contain points on both sides of M_p . Let $f(r)$ be a point of M furthest from M_p on the *other* side from q . Then $N(r)$ must equal $N(p)$, contradicting the fact that N is one-one. Thus we have shown that f is an embedding.

To show that $f(M)$ is convex, we use a similar argument. Given $p \in M$, we just have to show that all of $f(M)$ lies on the same side of $M_p \subset \mathbb{R}^3_{f(p)}$ as $N(p)$ does. If there were points on the opposite side, and $f(q)$ were a point on the opposite side which is furthest from M_p , then $N(q)$ would have to equal $N(p)$, again a contradiction. ♦

This result naturally invites comparison with Theorem II.1-8, which states that a simple closed curve c in \mathbb{R}^2 is convex if and only if it satisfies $\kappa \geq 0$ or $\kappa \leq 0$ (depending on the direction in which c is traversed). The proof of part (1) of Theorem 11 is much simpler than the proof of the corresponding part of Theorem II.1-8. This is because the sign of the Gaussian curvature K has a local geometric meaning, while the sign of κ has none; the only meaningful assertion about κ is the *global* statement that it is ≥ 0 or ≤ 0 everywhere. In part (2) of Theorem 11 we have the significant circumstance that we do not have to assume that M is imbedded—this comes out as part of the conclusion. The analogous assertion is *false* in the case of curves: the figure below shows an immersed, but not imbedded, curve with $\kappa \geq 0$ everywhere. In one respect



our result does not improve on Theorem II.1-8, for in order to prove part (2), we needed to assume the strict inequality $K(p) > 0$. Actually the result holds even when we assume only that $K \geq 0$, but the proof in this case is much more difficult (for further discussion, and references, see pp. IV. 82–83).

Our approach to surface theory has so far been very classical, but we are now ready to jazz it up a bit. First we want to examine the moving frame approach again, and write out explicitly all the equations (which in the case of surfaces in \mathbb{R}^3 boil down to almost nothing). In addition to their importance in the remainder of this chapter, some of these formulas will be crucial in Chapter 6 (and the equations in the general case will be even more crucial in Chapter 7).

If X_1, X_2 is a positively oriented orthonormal moving frame on an oriented surface $M \subset \mathbb{R}^3$, and we let $X_3 = \nu$, then X_1, X_2, X_3 is a positively oriented adapted orthonormal moving frame on M . There are just the following forms to consider:

$$\begin{array}{ll} \theta^1, \theta^2 & \text{the dual 1-forms} \\ \omega_1^2 = -\omega_2^1 & \text{the connection form} \\ \psi_1^3, \psi_2^3. & \end{array}$$

We want to relate these forms to the tensors and functions on M already considered. Notice first that \mathbf{I} is given by

$$\mathbf{I} = \theta^1 \otimes \theta^1 + \theta^2 \otimes \theta^2.$$

We also have

$$dA = \theta^1 \wedge \theta^2,$$

where dA is the volume element determined by the metric \mathbf{I} on M and the given orientation. Since we have (see page 19ff.)

$$\begin{aligned}\psi_1^3 &= s_{11}^3 \theta^1 + s_{21}^3 \theta^2 = \Pi(X_1, X_1) \theta^1 + \Pi(X_2, X_1) \theta^2 \\ \psi_2^3 &= s_{12}^3 \theta^1 + s_{22}^3 \theta^2 = \Pi(X_1, X_2) \theta^1 + \Pi(X_2, X_2) \theta^2,\end{aligned}$$

we can write

$$\Pi = \psi_1^3 \otimes \theta^1 + \psi_2^3 \otimes \theta^2.$$

It is also easy to see that the Gaussian and mean curvatures

$$\begin{aligned}K &= \Pi(X_1, X_1) \cdot \Pi(X_2, X_2) - [\Pi(X_1, X_2)]^2 \\ H &= \frac{1}{2} \{ \Pi(X_1, X_1) + \Pi(X_2, X_2) \}\end{aligned}$$

are given by

$$\begin{aligned}\psi_1^3 \wedge \psi_2^3 &= K \theta^1 \wedge \theta^2 \\ \psi_1^3 \wedge \theta^2 - \psi_2^3 \wedge \theta^1 &= 2H \theta^1 \wedge \theta^2.\end{aligned}$$

On the other hand, we have a much more important expression for K , in terms of the connection form ω_1^2 . We note first that equation (3) on page 16 now reduces to

$$d\omega_1^2 = \Omega_1^2.$$

Then Gauss' equation (page 20) becomes simply

$$0 = d\omega_1^2 - \psi_2^3 \wedge \psi_1^3 = d\omega_1^2 + K \theta^1 \wedge \theta^2,$$

so that

$$d\omega_1^2 = -K \theta^1 \wedge \theta^2.$$

Since this equation is equivalent to Gauss' equation, it must somehow demonstrate the Theorema Egregium, and it surely does, since the form ω_1^2 does not depend on the imbedding (it is the unique form with $d\theta^1 = \omega_1^2 \wedge \theta^2$ and $d\theta^2 = -\omega_1^2 \wedge \theta^1$). Some elementary treatments of surface theory proceed to

use this equation to *define* the Gaussian curvature of an arbitrary 2-dimensional Riemannian manifold—it is only necessary to check that K does not depend on the choice of the orthonormal moving frame; this is a special case of the moving frame definition of the curvature tensor given in Chapter II.7. Finally, we mention that the Codazzi-Mainardi equations (page 20) now become

$$\begin{aligned} d\psi_1^3 &= \omega_1^2 \wedge \psi_2^3 \\ d\psi_2^3 &= -\omega_1^2 \wedge \psi_1^3. \end{aligned}$$

An introduction to surface theory carried out purely in terms of this moving frame and structural equation approach can be very frustrating. Instead of dealing with geometrically tangible things like dN and Π , one has only the 1-forms $\omega_1^2, \psi_1^3, \psi_2^3$ to play with, and the simplicity of the Gauss and Codazzi-Mainardi equations as given above seems vitiated by their lack of intuitive geometric content. But this simplicity is a great advantage in proving theorems, and can be attributed, in large measure, to the fact that they express integrability conditions so neatly in terms of d . For example, they allow us to give a proof of the fundamental theorem of surface theory which uses the differential form version of the Frobenius integrability theorem (Proposition I.7-14), instead of mucking around with the classical integrability conditions; we will present this proof, in a more general situation, in Chapter 7. The truly overwhelming advantage of the moving frame approach becomes apparent when one is seriously investigating questions about the shape of surfaces in space; any information one can hope to get has to come out of the three simple equations

$$d\omega_1^2 = -\psi_1^3 \wedge \psi_2^3, \quad d\psi_1^3 = \omega_1^2 \wedge \psi_2^3, \quad d\psi_2^3 = -\omega_1^2 \wedge \psi_1^3.$$

Usually one just picks a frame suited to the problem and reads off the information from these equations. As a very simple example, we consider an all-umbilic surface $M \subset \mathbb{R}^3$. In this situation any adapted orthonormal moving frame X_1, X_2, X_3 on M is suitable (the hypothesis that p is an umbilic essentially says that all orthonormal frames at p are indistinguishable from one another), and we have

$$\psi_i^3 = \lambda \theta^i \quad i = 1, 2$$

for some function λ on M . Thus

$$\begin{aligned} d\lambda \wedge \theta^1 + \lambda d\theta^1 &= d\psi_1^3 = \omega_1^2 \wedge \psi_2^3 = \lambda \omega_1^2 \wedge \theta^2 \\ d\lambda \wedge \theta^2 + \lambda d\theta^2 &= -\lambda \omega_1^2 \wedge \theta^1, \end{aligned}$$

while

$$\begin{aligned} d\theta^1 &= \omega_1^2 \wedge \theta^2 \\ d\theta^2 &= -\omega_1^2 \wedge \theta^1. \end{aligned}$$

So we find that

$$d\lambda \wedge \theta^i = 0 \quad i = 1, 2.$$

But this implies that the 1-form $d\lambda$ is 0, so we find, once again, that λ is constant. More interesting examples will occur in later chapters.

The moving frame approach was brought in at this point not to launch an extended investigation into the geometry of surfaces—that occurs in later chapters—but with a completely different goal in mind. We want to show that the Gauss and Codazzi-Mainardi equations for surfaces in \mathbb{R}^3 are, from the proper point of view, nothing more than the “equations of structure” of the Lie group $SO(3)$, and that the Fundamental Theorem of Surface Theory reduces to Theorems I.10-17 and I.10-18 about Lie groups. After doing this, we will then proceed to bring another group into the picture by examining properties of surfaces in \mathbb{R}^3 which are invariant under the group of maps $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ of the form $A = T \circ B$ for T a translation and $B \in SL(3) =$ group of 3×3 matrices with $\det = 1$.

We begin with some preliminaries about notation. Nowadays, an “affine” map $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is usually defined to be one of the form $A = T \circ B$ for T a translation and $B \in GL(n, \mathbb{R})$. Thus the proper Euclidean motions, $A = T \circ B$ for $B \in SO(n)$, might be described as “special orthogonal affine” maps, while maps $A = T \circ B$ for $B \in SL(n)$ might best be described as “special linear affine” maps. We will employ this terminology regularly for maps $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$, but we will also find it convenient to abbreviate the phrase “special linear affine” to “special affine” when we speak of such concepts as “special affine curvature”. Thus when we speak of the “special affine geometry of surfaces” in \mathbb{R}^3 , we mean properties of surfaces invariant under special linear affine maps. On those occasions when we want to consider properties of surfaces invariant under all affine maps, we will emphasize this fact by speaking of “general affine” invariants.

We will also include a brief review of the relevant facts about Lie groups (pp. II.36–37), since we are going to make a slight change of notation. If G is a Lie group with Lie algebra \mathfrak{g} , then we define the natural \mathfrak{g} -valued 1-form* ω on G by $\omega(a)(\tilde{X}(a)) = X$, where \tilde{X} is the left invariant vector field with $\tilde{X}(e) = X$. If X_1, \dots, X_n is a basis of \mathfrak{g} , then we have $\omega = \sum_{i=1}^n \omega^i X_i$ for ordinary (\mathbb{R} -valued) left invariant 1-forms ω^i , and these forms ω^i are a basis for the left invariant 1-forms. These forms are important because two maps

*We are now using ω to distinguish this form on G from the forms ω_1^2 for a moving frame on a surface. This wasn't important in Volume II, where we considered only curves.

$f, g: M \rightarrow G$ differ by a left translation [$f = L_a \circ g$ for some $a \in G$] if and only if $f^*(\omega^i) = g^*(\omega^i)$ for all i (Theorem I.10-18). When G is a subgroup of $GL(n, \mathbb{R})$, with $P: G \rightarrow GL(n, \mathbb{R}) \subset \mathbb{R}^{n^2}$ the inclusion map, then we can form $P^{-1} \cdot dP$, where P^{-1} denotes (somewhat confusingly) the map $A \mapsto A^{-1}$, and the differential dP of P can be considered either as an \mathbb{R}^{n^2} -valued 1-form on G , or as the matrix of 1-forms $dP = (dx^{ij})$, where dx^{ij} denotes the differential of $x^{ij}|_G$. Then (pp. II.36–37) $P^{-1} \cdot dP$ is the natural \mathfrak{g} -valued 1-form ω on G . Among the entries of this matrix will be a basis for the left invariant 1-forms on G (the entries are generally not linearly independent, since the forms dx^{ij} are not linearly independent on G). So if $f: M \rightarrow GL(n, \mathbb{R})$ is a C^∞ map, and we want to look at the forms $f^*(\omega^i)$ for a basis $\{\omega^i\}$ of left invariant 1-forms on G , it suffices to look at the entries of the matrix

$$f^*(P^{-1} \cdot dP) = f^{-1} \cdot df.$$

To study properties of immersions $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ which are invariant under special orthogonal affine maps of \mathbb{R}^3 , we need to define an associated map $\alpha_f: \mathbb{R}^2 \rightarrow SO(3)$. This can be done in the following way. Let $X_3 = N$ be the normal map of f , and let X_1, X_2 be the result of applying the Gram-Schmidt orthonormalization process to the vectors f_1, f_2 . We then have an adapted orthonormal moving frame X_1, X_2, X_3 on \mathbb{R}^2 , and if we also consider X_i as a column vector, then $\alpha_f = (X_1, X_2, X_3): \mathbb{R}^2 \rightarrow SO(3)$ is the desired map. Notice that we can reconstruct f_1, f_2 from X_1, X_2 when we are also given the functions g_{ij} .

To find $\alpha_f^*(\omega^i)$, where the ω^i are a basis for the left invariant 1-forms on $SO(3)$, we look at the entries of the matrix of 1-forms

$$a = \alpha_f^{-1} \cdot d\alpha_f, \quad \text{with} \quad d\alpha_f = \alpha_f \cdot a.$$

This equation means that

$$(dX_1, dX_2, dX_3) = (X_1, X_2, X_3) \begin{pmatrix} 0 & -a_{21} & -a_{31} \\ a_{21} & 0 & -a_{32} \\ a_{31} & a_{32} & 0 \end{pmatrix},$$

where the a_{ij} are 1-forms, and the X_i and dX_i are considered as column vectors; the latter equation stands for

$$dX_1 = a_{21}X_2 + a_{31}X_3, \quad \text{etc.}$$

Thus, if ∇' denotes the covariant differentiation in \mathbb{R}^3 , we have

$$\nabla'_X X_1 = dX_1(X) = a_{21}(X)X_2 + a_{31}(X)X_3, \quad \text{etc.}$$

But

$$\nabla'_X X_1 = \omega_1^2(X) \cdot X_2 + \psi_1^3(X) \cdot X_3, \quad \text{etc.}$$

(where $\omega_1^2, \psi_1^3, \psi_2^3$ really denote f^* of the corresponding forms on image f). Thus we see that

The forms $\omega_1^2, \psi_1^3, \psi_2^3$ are precisely $\alpha_f^(\omega^i)$ for ω^i a basis of the left invariant 1-forms on $\text{SO}(3)$.*

Theorem I.10-18 then tells us that for two immersions $f, \bar{f}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$, the maps $\alpha_f, \alpha_{\bar{f}}: \mathbb{R}^2 \rightarrow \text{SO}(3)$ differ by an element of $\text{SO}(3)$ if and only if

$$\bar{\omega}_1^2 = \omega_1^2, \quad \bar{\psi}_1^3 = \psi_1^3, \quad \bar{\psi}_2^3 = \psi_2^3;$$

here the forms $\omega_1^2, \psi_1^3, \psi_2^3$ are formed for the moving frame $X_1, X_2, X_3 = N$, where X_1, X_2 are obtained by applying the Gram-Schmidt orthonormalization process to f_1, f_2 , while the forms $\bar{\omega}_1^2, \bar{\psi}_1^3, \bar{\psi}_2^3$ are formed for the moving frame $\bar{X}_1, \bar{X}_2, \bar{X}_3 = \bar{N}$, where \bar{X}_1, \bar{X}_2 are obtained by applying the Gram-Schmidt orthonormalization process to \bar{f}_1, \bar{f}_2 .

From this fact we can easily derive the first part of the Fundamental Theorem of Surface Theory. For if $\bar{g}_{ij} = g_{ij}$, then \bar{X}_1, \bar{X}_2 are the same linear combination of \bar{f}_1, \bar{f}_2 as X_1, X_2 are of f_1, f_2 . Consequently, if we are also given that $\bar{l}_{ij} = l_{ij} \implies \bar{\Pi}(\bar{f}_i, \bar{f}_j) = \Pi(f_i, f_j)$, then we conclude that $\bar{\Pi}(\bar{X}_i, \bar{X}_j) = \Pi(X_i, X_j)$. The formulas on page 69 then show that $\bar{\psi}_1^3 = \psi_1^3$ and $\bar{\psi}_2^3 = \psi_2^3$, while the equation $\bar{\omega}_1^2 = \omega_1^2$ follows from $\bar{g}_{ij} = g_{ij}$. Thus $\alpha_f, \alpha_{\bar{f}}: \mathbb{R}^2 \rightarrow \text{SO}(3)$ differ by an element of $\text{SO}(3)$; this implies that $(f_1, f_2, N), (\bar{f}_1, \bar{f}_2, \bar{N})$ differ by an element of $\text{SO}(3)$, and hence that f, \bar{f} differ by a special orthogonal affine map of \mathbb{R}^3 .

In the case of curves, the equations of structure of $\text{SO}(n)$ or $\text{SL}(n, \mathbb{R})$ could not give any interesting information, since there are no non-zero 2-forms on \mathbb{R} . But they do give information for surfaces. To figure out the equations of structure of $\text{SO}(3)$, we proceed as follows. Since the Lie algebra

$$\mathfrak{o}(3) = \{\text{tangent space of } \text{SO}(3) \text{ at } I\}$$

is just the set of skew-symmetric 3×3 matrices, the matrices

$$X_1 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

are a basis for $\mathfrak{o}(3)$. The bracket operation in $\mathfrak{o}(3)$ is (pp. I.378–379)

$$[M, N] = MN - NM.$$

In particular, we compute that

$$[X_1, X_2] = X_3, \quad [X_1, X_3] = -X_2, \quad [X_2, X_3] = X_1,$$

so that if we write

$$[X_i, X_j] = \sum_{k=1}^3 C_{ij}^k X_k,$$

then the “constants of structure” C_{ij}^k are

$$\begin{array}{lll} C_{12}^1 = 0 & C_{12}^2 = 0 & C_{12}^3 = 1 \\ C_{13}^1 = 0 & C_{13}^2 = -1 & C_{13}^3 = 0 \\ C_{23}^1 = 1 & C_{23}^2 = 0 & C_{23}^3 = 0. \end{array}$$

Now let ω^i be the left invariant 1-forms on $\mathfrak{o}(3)$ with $\omega^1(I)$, $\omega^2(I)$, $\omega^3(I)$ dual to X_1 , X_2 , X_3 . The equation on pg. I.396 (which is equivalent to the “equations of structure” on pg. I.404) then gives

$$\begin{aligned} d\omega^1 &= -\omega^2 \wedge \omega^3 \\ d\omega^2 &= \omega^1 \wedge \omega^3 \\ d\omega^3 &= -\omega^1 \wedge \omega^2. \end{aligned}$$

Now we have seen that if $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is an immersion, then the forms ω_1^2 , ψ_1^3 , ψ_2^3 for f are given by

$$\begin{aligned} \omega_1^2 &= \alpha_f^*(\omega^1) \\ \psi_1^3 &= \alpha_f^*(\omega^2) \\ \psi_2^3 &= \alpha_f^*(\omega^3). \end{aligned}$$

Therefore

$$\begin{aligned} d\omega_1^2 &= \alpha_f^*(d\omega^1) = -\alpha_f^*(\omega^2 \wedge \omega^3) = -\psi_1^3 \wedge \psi_2^3 \\ d\psi_1^3 &= \alpha_f^*(d\omega^2) = -\alpha_f^*(\omega^1 \wedge \omega^3) = \omega_1^2 \wedge \psi_2^3 \\ d\psi_2^3 &= \alpha_f^*(d\omega^3) = -\alpha_f^*(\omega^1 \wedge \omega^2) = -\omega_1^2 \wedge \psi_1^3. \end{aligned}$$

As promised, these are precisely the Gauss Equation and Codazzi-Mainardi Equations, in the form given on pages 69–70. The reader can now easily see that the second part of the Fundamental Theorem of Surface Theory follows immediately from Theorem I.10-17.

For the remainder of this chapter* we will be considering the special affine theory of surfaces $M \subset \mathbb{R}^3$. If we try to follow the approach used for ordinary surface theory, then instead of working with an adapted moving frame X_1, X_2, ν on M which is orthonormal, we want to work with an adapted moving frame X_1, X_2, X_3 on M with $\det(X_1, X_2, X_3) = 1$. If $f: U \rightarrow M$ is an immersion (for $U \subset \mathbb{R}^2$ open), then you might think that we should use the moving frame

$$(af_1, af_2, aN), \quad \text{where} \quad a = \frac{1}{\sqrt[3]{\det(f_1, f_2, N)}}.$$

But this moving frame has no significance for special affine geometry, for it is not a “special affine invariant”: if $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is special linear affine, then the normal $N_{A \circ f}$ for $A \circ f$ is not necessarily the image $A_*(N_f)$ of the normal N_f for f . As a matter of fact, not only the length, but even the direction of $A_*(N_f)$ will be wrong; the whole concept of “orthogonality” has no meaning in special affine geometry. Our first problem, therefore, is to pick out a “special affine normal” for M which is a special affine invariant. This is going to take quite a bit of doing.

In ordinary surface theory, the normal $\nu(p)$ is defined in terms of the tangent plane M_p of M , which is the first order surface which approximates M up to order 1 at p . Since special affine geometry always seems to involve higher order approximations to our given geometric object, we might expect to find a reasonable candidate for the special affine normal by looking at the second order approximation to our surface. As before, let us assume that $p = 0 \in \mathbb{R}^3$ and that the tangent plane at p is the (x, y) -plane, so that M is the graph of a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ with $h(0, 0) = h_1(0, 0) = h_2(0, 0) = 0$. The quadratic surface approximating M up to order 2 at 0 is

$$P = \left\{ (s, t, \frac{1}{2}(h_{11}(0, 0) \cdot s^2 + 2h_{12}(0, 0) \cdot st + h_{22}(0, 0) \cdot t^2)) \right\}.$$

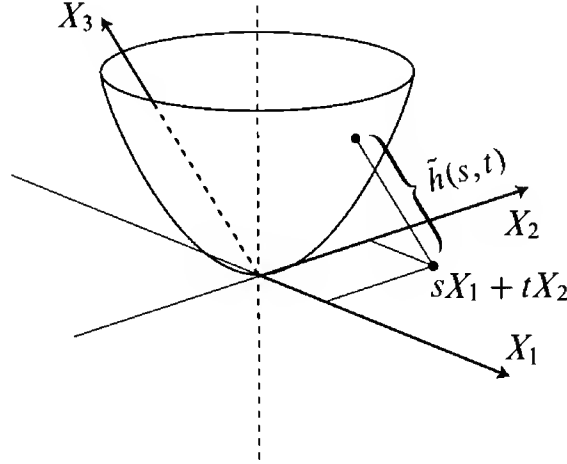
We have already seen that this surface does not depend on the particular choice of a basis in \mathbb{R}^2 : if $\mathbf{X} = X_1, X_2$ is any basis of \mathbb{R}^2 and $h^{\mathbf{X}}(s, t)$ is the third coordinate of the point of M lying above $sX_1 + tX_2$, then the surface

$$Q = \left\{ (sX_1 + tX_2, \frac{1}{2}(h^{\mathbf{X}}_{11}(0, 0) \cdot s^2 + 2h^{\mathbf{X}}_{12}(0, 0) \cdot st + h^{\mathbf{X}}_{22}(0, 0) \cdot t^2)) \right\}$$

is exactly the same as P . We also noted that we still have $Q = P$ when we change the direction of the z -axis; it is just as easy to see that $Q = P$ even when we change the *unit* on the z -axis, so that we are describing M in terms of the “ $X_1, X_2, (0, 0, c)$ coordinate system”.

*This material will not be needed later, except in Problem 4-16. Problems for this chapter are on page 134.

In all this arbitrariness, however, one essential prejudice of ordinary geometry remains: we have always picked the third axis perpendicular to the plane of the first two. Suppose now that we choose X_1, X_2 in \mathbb{R}^2 and a linearly independent vector X_3 which does not necessarily point along the z -axis. We can



still describe M as a graph in terms of the “ (X_1, X_2, X_3) coordinate system”: we let $\tilde{h}(s, t)$ be the X_3 component of the point of M with $s = X_1$ component and $t = X_2$ component. Now we look at the surface

$$Q = \{sX_1 + tX_2 + \frac{1}{2}(\tilde{h}_{11}(0, 0) \cdot s^2 + 2\tilde{h}_{12}(0, 0) \cdot st + \tilde{h}_{22}(0, 0) \cdot t^2)X_3\}.$$

This surface is *not* the same surface as P . In fact, consider the case where $X_3 = (0, 0, 1) + \lambda X_1 + \mu X_2$ for certain numbers λ, μ . To say that M is the graph of h in the $X_1, X_2, (0, 0, 1)$ system means that

$$(1) \quad M = \{sX_1 + tX_2 + h(s, t) \cdot (0, 0, 1)\} \quad [h_\alpha(0, 0) = 0];$$

similarly, if M is the graph of \tilde{h} in the X_1, X_2, X_3 system, then

$$(2) \quad \begin{aligned} M &= \{sX_1 + tX_2 + \tilde{h}(s, t)X_3\} \\ &= \{[s + \lambda\tilde{h}(s, t)]X_1 + [t + \mu\tilde{h}(s, t)]X_2 + \tilde{h}(s, t) \cdot (0, 0, 1)\} \\ &\quad [\tilde{h}_\alpha(0, 0) = 0]. \end{aligned}$$

Comparing (1) and (2) we find that

$$\tilde{h}(s, t) = h(s + \lambda\tilde{h}(s, t), t + \mu\tilde{h}(s, t)).$$

From this we easily compute that $\tilde{h}_{\alpha\beta}(0, 0) = h_{\alpha\beta}(0, 0)$, so that the approximating functions of s and t

$$\begin{aligned} &\tilde{h}_{11}(0, 0)s^2 + 2\tilde{h}_{12}(0, 0)st + \tilde{h}_{22}(0, 0)t^2 \\ &h_{11}(0, 0)s^2 + 2h_{12}(0, 0)st + h_{22}(0, 0)t^2 \end{aligned}$$

are the same; this means that the approximating *surfaces*

$$\begin{aligned} P &= \{(sX_1 + tX_2, \tfrac{1}{2}(h_{11}(0,0)s^2 + 2h_{12}(0,0)st + h_{22}(0,0)t^2))\} \\ Q &= \{sX_1 + tX_2 + \tfrac{1}{2}(\tilde{h}_{11}(0,0)s^2 + 2\tilde{h}_{12}(0,0)st + \tilde{h}_{22}(0,0)t^2)X_3\} \\ &= \{sX_1 + tX_2 + \tfrac{1}{2}(h_{11}(0,0)s^2 + 2h_{12}(0,0)st + h_{22}(0,0)t^2)X_3\} \end{aligned}$$

are definitely different. In fact, we clearly have

$$Q = A(P),$$

where $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is the affine map which keeps M_p fixed and takes $(0, 0, 1)$ to X_3 . All we can say is that Q does not change when we multiply X_3 by a constant, just as P does not change when we multiply $(0, 0, 1)$ by a constant; we can merely speak of the osculating paraboloid corresponding to any given line through p which does not lie in M_p .

Thus we see that we do not get a special affine invariant osculating paraboloid simply by looking at M up to order 2. There are some things that we do get, however. Consider first a fixed basis X_1, X_2 for $M_p = (x, y)$ -plane. We have just seen that the matrices

$$S = \begin{pmatrix} h_{11}(0,0) & h_{12}(0,0) \\ h_{21}(0,0) & h_{22}(0,0) \end{pmatrix} \quad \text{and} \quad S' = \begin{pmatrix} \tilde{h}_{11}(0,0) & \tilde{h}_{12}(0,0) \\ \tilde{h}_{21}(0,0) & \tilde{h}_{22}(0,0) \end{pmatrix},$$

defined in terms of the third axes $(0, 0, 1)$ and $X_3 = (0, 0, 1) + \lambda X_1 + \mu X_2$, respectively, are exactly the same; if we had picked X_3 to be the most general possible choice, $X_3 = (0, 0, c) + \lambda X_1 + \mu X_2$, then S' would clearly be

$$S' = \frac{1}{c} \cdot S.$$

On the other hand, suppose we consider the coordinate system

$$a_{11}X_1 + a_{21}X_2, \quad a_{12}X_1 + a_{22}X_2, \quad (0, 0, 1).$$

Equation (*) on page 37 shows that the matrix S' in this case is related to the matrix S by

$$S' = A^t S A, \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

where t denotes the transpose. In general, if we are given any two bases (X_1, X_2, X_3) and (X'_1, X'_2, X'_3) of \mathbb{R}^3_p , with X_1, X_2 and X'_1, X'_2 bases for M_p , and

$$X'_3 = cX_3 + \lambda X_1 + \mu X_2,$$

then the matrices S and S' are related by

$$(1) \quad S' = \frac{1}{c} \cdot A^t S A,$$

where $A = (a_{ij})$ is given by

$$(2) \quad X'_i = \sum_{j=1}^2 a_{ji} X_j.$$

From equation (1) we see that

$$\det S' \gtrless 0 \iff \det S \gtrless 0 \quad \text{and} \quad S' = 0 \iff S = 0.$$

Thus we can determine whether p is an elliptic, hyperbolic, parabolic, or planar, point of M by means of an arbitrary basis (X_1, X_2, X_3) of \mathbb{R}^3_p with $X_1, X_2 \in M_p$. This clearly implies that the “type” of a point (elliptic, hyperbolic, parabolic, or planar) is a “general affine invariant”: If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is any affine map and $M \subset \mathbb{R}^3$ is a surface, then $A(p)$ is the same type of point on $A(M)$ as p is on M .

Consider, for the time being, an elliptic point $p \in M$. As we observed in the proof of Theorem 11, there is a natural orientation for M_p , the one that makes an ordered basis (X_1, X_2) of M_p positively oriented whenever (X_1, X_2, X_3) is positively oriented in \mathbb{R}^3_p for any $X_3 \in \mathbb{R}^3_p$ which points “inward” (that is, in the direction of the osculating paraboloid). Now any basis $X_1, X_2, X_3 \in \mathbb{R}^3_p$ with $X_1, X_2 \in M_p$ determines a matrix $S = (s_{ij})$, and we can use S to define an inner product on M_p by

$$\langle X_i, X_j \rangle = s_{ij};$$

this inner product is positive definite if and only if X_3 is inward pointing. If X_1, X_2, X_3 happens to be orthonormal, then the inner product is just the second fundamental form II of ordinary surface theory. Now consider another basis X'_1, X'_2, X'_3 , with X'_3 inward pointing. This determines a matrix S' , and hence another positive definite inner product $\langle \cdot, \cdot \rangle'$ on M_p by

$$\langle X'_i, X'_j \rangle' = s'_{ij}.$$

Equation (1) shows that the inner products $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle'$ are closely related. Indeed,

$$\begin{aligned} \langle X'_i, X'_j \rangle' &= s'_{ij} = (\text{constant}) \cdot \sum_{k,l} a_{ki} s_{kl} a_{lj} \\ &= (\text{constant}) \cdot \sum_{k,l} a_{ki} a_{lj} \langle X_k, X_l \rangle \\ &= (\text{constant}) \cdot \left\langle \sum_k a_{ki} X_k, \sum_l a_{lj} X_l \right\rangle \\ &= (\text{constant}) \cdot \langle X'_i, X'_j \rangle \quad \text{by (2),} \end{aligned}$$

so that we have

$$\langle \cdot, \cdot \rangle' = (\text{constant}) \cdot \langle \cdot, \cdot \rangle.$$

By restricting our attention to inward pointing X_3 , we thus obtain a class of positive definite inner products on M_p , any one being a (positive) constant multiple of any other. We can express this by saying that we have defined a “conformal structure” on M_p (compare pg. II.296). This conformal structure is invariant under all affine maps, provided only that they are orientation preserving: If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is any orientation preserving affine map and $p \in M \subset \mathbb{R}^3$ is an elliptic point, then the class of inner products defined on M_p is precisely A^* of the class of inner products defined on the tangent space $A(M)_{A(p)}$ of $A(M)$ at $A(p)$.

A conformal structure on M_p does not allow us to pick out orthonormal bases, but it does make sense to consider bases X_1, X_2 of M_p with

$$\langle X_i, X_j \rangle = (\text{constant} > 0) \cdot \delta_{ij},$$

since this condition does not depend on the choice of the inner product $\langle \cdot, \cdot \rangle$ in our conformal structure. Such bases may be called “quasi-orthonormal”. They can clearly be characterized quite simply as follows: for any inward pointing vector $X_3 \in \mathbb{R}^3_p$, the corresponding osculating paraboloid P is given by

$$P = \{sX_1 + tX_2 + (\text{constant} > 0) \cdot (s^2 + t^2) \cdot X_3\}.$$

These considerations can be made in a less geometric way, but with the calculations going through more smoothly, by using moving frames. For an adapted moving frame X_1, X_2, X_3 on a surface $M \subset \mathbb{R}^3$ we will still use the dual and connection forms ϕ^α and ψ^α_β for moving frames in \mathbb{R}^3 , and we again let θ^1, θ^2

be the dual forms determined by the moving frame X_1, X_2 on M . As in ordinary surface theory, we have

$$\begin{aligned}\phi^i &= \theta^i & \text{on } TM & \quad i = 1, 2 \\ \phi^3 &= 0 & \text{on } TM;\end{aligned}$$

moreover the first structural equation,

$$d\phi^3 = - \sum_{\gamma=1}^3 \psi_\gamma^3 \wedge \phi^\gamma,$$

still implies that

$$0 = \sum_{k=1}^2 \theta^k \wedge \psi_k^3 \quad \text{on } TM,$$

so that by Cartan's Lemma there is a matrix $S = (s_{ij})$ with

$$\begin{aligned}\text{(a)} \quad \psi_j^3 &= \sum_{i=1}^2 s_{ij} \theta^i & \text{on } TM \\ s_{ij} &= s_{ji}.\end{aligned}$$

We will soon be able to compare this matrix S with the one defined previously. First we want to consider another adapted moving frame X'_1, X'_2, X'_3 on M . The matrix a with $X'_\alpha = \sum_\beta a_{\beta\alpha} X_\beta$ must be of the form

$$a = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix},$$

and we easily find that

$$\text{(b)} \quad (a^{-1})_{3i} = 0 \quad i = 1, 2; \quad (a^{-1})_{33} = \frac{1}{a_{33}}.$$

The dual forms ϕ'^α for the X'_α are given by (see pg. II. 282)

$$\begin{aligned}\text{(c)} \quad \phi' &= a^{-1} \phi \implies \phi = a \cdot \phi' \\ &\implies \theta^i = \sum_{j=1}^2 a_{ij} \theta'^j,\end{aligned}$$

while the connection forms ψ'^α_β are related to the ψ^α_β by (pg. II.280)

$$(d) \quad \psi' = a^{-1} da + a^{-1} \psi a.$$

In particular,

$$\begin{aligned} \psi'^3_i &= (a^{-1} da)_{3i} + (a^{-1} \psi a)_{3i} \\ &= \sum_{\alpha=1}^3 (a^{-1})_{3\alpha} da_{\alpha i} + \sum_{\alpha, \beta=1}^3 (a^{-1})_{3\alpha} \psi^\alpha_\beta a_{\beta i} \\ &= 0 + \sum_{j=1}^2 \frac{1}{a_{33}} \psi^3_j a_{ji} \quad \text{by (b).} \end{aligned}$$

Hence

$$\begin{aligned} \psi'^3_i &= \frac{1}{a_{33}} \sum_{j,k} a_{ji} s_{jk} \theta^k \\ &= \frac{1}{a_{33}} \sum_{j,k,l} a_{ji} s_{jk} a_{kl} \theta'^l. \end{aligned}$$

So if we also write

$$\psi'^3_j = \sum_{i=1}^2 s'_{ij} \theta'^i, \quad s'_{ij} = s'_{ji},$$

then the matrix S' is related to the matrix S by

$$(e) \quad S' = \frac{1}{a_{33}} A^t S A,$$

where A is the 2×2 matrix $A = (a_{ij})$.

As a first consequence of this equation we see, what is not *a priori* clear, that

The matrix $S(p)$ depends only on the vectors $X_1(p), X_2(p), X_3(p)$.

Taking X_3 to be a parallel vector field in \mathbb{R}^3 , we easily see that $S(p)$ is, in fact, the same as the matrix S on page 77, for the basis $X_1(p), X_2(p), X_3(p)$ of \mathbb{R}^3_p . Then equation (e) is just equation (l) on page 78. As before, we then see from equation (e) that if p is an elliptic point, then the inner product

$$\sum_{i,j} s_{ij}(p) \cdot \theta^i(p) \otimes \theta^j(p)$$

on M_p is well-defined up to constant multiple, and hence, by considering only the case where $X_3(p)$ is inward pointing, we again obtain the (general orientation preserving affine invariant) conformal structure on M_p . Clearly the basis X_1, X_2 of M_p is quasi-orthonormal [that is, $\langle X_i(p), X_j(p) \rangle = (\text{constant} > 0) \cdot \delta_{ij}$] if and only if

$$S(p) = (\text{constant} > 0) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \psi_i^3(p) = (\text{constant} > 0) \cdot \theta^i(p).$$

Now from among our class of inner products on M_p we can distinguish a particular one $\langle \cdot, \cdot \rangle_p$, which will be a special affine invariant. We do this by defining $X_1, X_2 \in M_p$ to be **orthonormal with respect to** $\langle \cdot, \cdot \rangle_p$ if and only if

$$(*) \quad S(p) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{for all inward pointing } X_3 \in \mathbb{R}^3_p \text{ with } \det(X_1, X_2, X_3) = \pm 1$$

(the sign depending on whether or not (X_1, X_2) is positively oriented). To check that this is well-defined, note first that if we also have

$$\det(X_1, X_2, X'_3) = \pm 1$$

for an inward pointing X'_3 , then clearly $a_{33}(p) = 1$, so equation (e) shows that $S'(p) = S(p)$; consequently, condition (*) does not depend on the choice of X_3 . Moreover, for fixed $X_3 \in \mathbb{R}^3_p$, and different $X'_1, X'_2 \in M_p$, equation (e) shows that $S' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ if and only if $A^t A = I$, which means that X'_1, X'_2 is related to X_1, X_2 by the orthogonal transformation A ; hence the inner product which makes X_1, X_2 orthonormal also makes X'_1, X'_2 orthonormal. It should be clear, from the very definition, that $\langle \cdot, \cdot \rangle_p$ is a special affine invariant: If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is special linear affine, and $p \in M$ is an elliptic point, then the inner product $\langle \cdot, \cdot \rangle_p$ on M_p is A^* of the inner product $\langle \cdot, \cdot \rangle_{A(p)}$ on the tangent space $A(M)_{A(p)}$ of $A(M)$ at $A(p)$. Clearly a basis $X_1, X_2 \in M_p$ is orthonormal with respect to $\langle \cdot, \cdot \rangle_p$ if and only if

$$\psi_i^3(p) = \theta^i(p) \quad \text{or} \quad P = \{sX_1 + tX_2 + \frac{1}{2}(s^2 + t^2)X_3\}$$

for every inward pointing $X_3 \in \mathbb{R}^3_p$ with $\det(X_1, X_2, X_3) = \pm 1$.

On a surface M with all points elliptic, we now have an inner product $\langle \cdot, \cdot \rangle_p$ defined on each M_p , and thus we have a Riemannian metric $\langle \cdot, \cdot \rangle$ on M . This Riemannian metric will also be denoted by \mathbb{I} , and called the **special affine first fundamental form** of M . If it seems strange that we can find a special affine invariant metric on elliptically curved surfaces in \mathbb{R}^3 even though there is clearly

no special affine invariant metric on \mathbb{R}^3 itself, it might help to observe that we have essentially the same situation for 1-dimensional manifolds M in \mathbb{R}^3 , for we can define the unit tangent vectors of M to be those of the form $c'(\sigma)$, where $c: [0, 1] \rightarrow M$ is a curve parameterized by special affine arclength. Once again the manifold M cannot be too flat (and in fact the requisite condition is more stringent, since it involves third derivatives). Naturally, if $f: M \rightarrow \mathbb{R}^3$ is an immersion with all points of the image elliptic, then we define the **special affine first fundamental form** \mathcal{I}_f of f to be the tensor $\mathcal{I}_f = f^*(\langle \cdot, \cdot \rangle)$ on M , where $\langle \cdot, \cdot \rangle$ is the special affine first fundamental form on $f(M)$. Notice that this is not completely analogous to the definition in ordinary surface theory, where we can simply define $\mathcal{I}_f = f^*\langle \cdot, \cdot \rangle$ for $\langle \cdot, \cdot \rangle$ the usual Riemannian metric on \mathbb{R}^3 ; it is much closer to the definition of \mathcal{II}_f . In fact, we have already noted that $\langle \cdot, \cdot \rangle_p$ is a multiple of $\mathcal{II}(p)$; just which multiple will soon be determined.

When $f: U \rightarrow \mathbb{R}^3$ for $U \subset \mathbb{R}^2$ open, and all points of $f(U)$ are elliptic, we define the functions $g_{ij}: U \rightarrow \mathbb{R}$ to be the components of \mathcal{I}_f with respect to the standard coordinate system (s, t) on \mathbb{R}^2 , so that

$$\mathcal{I}_f = g_{11} ds \otimes ds + g_{12} ds \otimes dt + g_{21} dt \otimes ds + g_{22} dt \otimes dt.$$

We would naturally like to be able to compute the g_{ij} in terms of f . We first take the case where f is simply

$$f(s, t) = (s, t, h(s, t)), \quad h(0, 0) = h_1(0, 0) = h_2(0, 0) = 0,$$

with $p = f(0, 0) = 0 \in \mathbb{R}^3$. If $X_1, X_2, X_3 \in \mathbb{R}^3_0$ is the standard basis, then the corresponding osculating paraboloid P is the graph of

$$(s, t) \mapsto \frac{1}{2}(\alpha s^2 + 2\beta st + \gamma t^2) \quad \text{for} \quad \begin{cases} \alpha = h_{11}(0, 0) \\ \beta = h_{12}(0, 0) \\ \gamma = h_{22}(0, 0). \end{cases}$$

If p is an elliptic point, then $\alpha\gamma - \beta^2 > 0$. For the sake of concreteness, suppose also that P lies above the (x, y) -plane, so that X_3 is inward pointing. There is another basis $X'_j = \sum_{i=1}^2 a_{ij} X_i$ of \mathbb{R}^2 such that P is the graph of

$$(s, t) \mapsto \frac{1}{2}(s^2 + t^2)$$

in the X'_1, X'_2, X_3 system. Equation (l) on page 78 (or equation (e) on page 81) shows that the matrix $A = (a_{ij})$ satisfies

$$\begin{aligned} (1) \quad A^t \cdot \frac{1}{2} \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \cdot A &= \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \implies \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} = (A^t)^{-1} A^{-1} \\ &\implies \det A = (\alpha\gamma - \beta^2)^{-1/2}. \end{aligned}$$

Now P is also the graph of $(s, t) \mapsto \frac{1}{2}(s^2 + t^2)$ in the $(\lambda X'_1, \lambda X'_2, \lambda^2 X_3)$ coordinate system. In order to have

$$\begin{aligned} 1 &= \det(\lambda X'_1, \lambda X'_2, \lambda^2 X_3) = \lambda^4 \det(X'_1, X'_2, X_3) \\ &= \lambda^4 \cdot \det A = \lambda^4 (\alpha\gamma - \beta^2)^{-1/2}, \end{aligned}$$

we must take

$$\lambda = \sqrt[8]{\alpha\gamma - \beta^2}.$$

So the vectors

$$X''_1 = (\sqrt[8]{\alpha\gamma - \beta^2}) X'_1, \quad X''_2 = (\sqrt[8]{\alpha\gamma - \beta^2}) X'_2$$

are orthonormal with respect to $\langle \cdot, \cdot \rangle_0$. To figure out the numbers $g_{ij}(0, 0)$, we note that if $B = (b_{ij})$ is the inverse of the matrix A , then

$$X_i = \sum_{j=1}^2 b_{ji} X'_j = \frac{1}{\sqrt[8]{\alpha\gamma - \beta^2}} \sum_{j=1}^2 b_{ji} X''_j,$$

and consequently

$$\begin{aligned} g_{ij}(0, 0) &= \langle X_i, X_j \rangle_0 = \frac{1}{\sqrt[4]{\alpha\gamma - \beta^2}} \left\langle \sum_{k=1}^2 b_{ki} X''_k, \sum_{k=1}^2 b_{kj} X''_k \right\rangle_0 \\ &= \frac{1}{\sqrt[4]{\alpha\gamma - \beta^2}} \sum_{k=1}^2 b_{ki} b_{kj} \\ &= \frac{1}{\sqrt[4]{\alpha\gamma - \beta^2}} (B^t B)_{ij} = \frac{1}{\sqrt[4]{\alpha\gamma - \beta^2}} ((A^t)^{-1} A^{-1})_{ij}. \end{aligned}$$

Using (1), we see that

$$(2) \quad g_{ij}(0, 0) = \frac{h_{ij}(0, 0)}{\sqrt[4]{\det(h_{ij}(0, 0))}}.$$

This can also be written

$$(3) \quad g_{ij}(0, 0) = \frac{d_{ij}}{\sqrt[4]{\det(d_{ij})}}(0, 0),$$

where $d_{ij} = \det(f_1, f_2, f_{ij})$.

These calculations were all carried out for the case where f is of the form $f(s, t) = (s, t, h(s, t))$, with $h(0, 0) = h_1(0, 0) = h_2(0, 0) = 0$. We could try to

deal with the general map $f: U \rightarrow \mathbb{R}^3$ by reducing it to this case. For example, if $f_1(0,0)$ and $f_2(0,0)$ both lie in the (x, y) -plane, we could first determine a function h by the condition

$$\begin{aligned} \{(f^1(s, t), f^2(s, t), f^3(s, t))\} &= f(U) = \{(s, t, h(s, t))\} \\ \implies h(f^1(s, t), f^2(s, t)) &= f^3(s, t), \end{aligned}$$

use this equation to relate the $h_{ij}(0,0)$ to the $f_i(0,0)$ and $f_{ij}(0,0)$, and then use (2) to find $g_{ij}(0,0)$ in terms of these numbers (we would still have to take care of the general case when $f_1(0,0)$ and $f_2(0,0)$ do not lie in \mathbb{R}^2). A much simpler course of action is to guess from (3) that the answer should be

$$\begin{aligned} \mathfrak{I}_f &= \frac{d_{11} ds \otimes ds + d_{12} ds \otimes dt + d_{21} dt \otimes ds + d_{22} dt \otimes dt}{\sqrt[4]{\det(d_{ij})}} \\ &= \frac{1}{\sqrt[4]{\det(d_{ij})}} \sum_{i,j=1}^2 d_{ij} ds^i \otimes ds^j, \quad \text{using } (s^1, s^2) \text{ for } (s, t). \end{aligned}$$

Now this guess cannot be precisely correct, for if we define $\tilde{f}(s^1, s^2) = f(s^2, s^1)$, then we have $\tilde{d}_{ij}(s^1, s^2) = \det(\tilde{f}_1, \tilde{f}_2, \tilde{f}_{12})(s^1, s^2) = \det(f_2, f_1, f_{12})(s^2, s^1) = -d_{ij}(s^2, s^1)$, so our formula changes sign. The problem, of course, is that \tilde{f} is orientation reversing if f is orientation preserving. The right guess is that the above formula holds whenever $f: U \rightarrow M$ is orientation preserving. To prove this, we note first that the right side is clearly a special affine invariant, since it involves only determinants d_{ij} ; consequently, there is no loss of generality in assuming that the tangent plane at the point in question is the (x, y) -plane. We still based our calculations on a very special parameterization, so we want to check that the right side is “invariant under orientation preserving change of parameter”: if $f: U \rightarrow \mathbb{R}^3$ is any immersion (for $U \subset \mathbb{R}^2$ open), and $p = (p^1, p^2): V \rightarrow U$ is an orientation preserving diffeomorphism (for $V \subset \mathbb{R}^2$ open), then we want to check that

$$(4) \quad p^* \left(\frac{1}{\sqrt[4]{\det(d_{ij})}} \sum_{i,j=1}^2 d_{ij} ds^i \otimes ds^j \right) = \frac{1}{\sqrt[4]{\det(\tilde{d}_{ij})}} \sum_{i,j=1}^2 \tilde{d}_{ij} ds^i \otimes ds^j,$$

where the \tilde{d}_{ij} are the d_{ij} for $\tilde{f} = f \circ p$.

Now

$$\begin{aligned}
 (5) \quad p^* \left(\sum_{i,j=1}^2 d_{ij} ds^i \otimes ds^j \right) &= \sum_{i,j=1}^2 (d_{ij} \circ p) \left(\sum_{\rho=1}^2 p^i_{\rho} ds^{\rho} \right) \otimes \left(\sum_{\sigma=1}^2 p^j_{\sigma} ds^{\sigma} \right) \\
 &= \sum_{i,j=1}^2 \sum_{\rho,\sigma=1}^2 p^i_{\rho} p^j_{\sigma} (d_{ij} \circ p) ds^{\rho} \otimes ds^{\sigma} \\
 &= \sum_{i,j=1}^2 \left(\sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (d_{\rho\sigma} \circ p) \right) ds^i \otimes ds^j.
 \end{aligned}$$

On the other hand, since

$$\begin{aligned}
 \tilde{f}_i &= D_i(f \circ p) = \sum_{\rho=1}^2 p^{\rho}_i (f_{\rho} \circ p) \\
 \tilde{f}_{ij} &= \sum_{\rho=1}^2 p^{\rho}_{ij} (f_{\rho} \circ p) + \sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (f_{\rho\sigma} \circ p),
 \end{aligned}$$

we have

$$\begin{aligned}
 (6) \quad \tilde{d}_{ij} &= \det(\tilde{f}_1, \tilde{f}_2, \tilde{f}_{ij}) \\
 &= \det \left(\sum_{\mu=1}^2 (f_{\mu} \circ p) \cdot p^{\mu}_1, \sum_{v=1}^2 (f_v \circ p) \cdot p^v_2, \right. \\
 &\quad \left. \sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (f_{\rho\sigma} \circ p) + \sum_{\rho=1}^2 p^{\rho}_{ij} (f_{\rho} \circ p) \right) \\
 &= \det \left(\sum_{\mu=1}^2 (f_{\mu} \circ p) \cdot p^{\mu}_1, \sum_{v=1}^2 (f_v \circ p) \cdot p^v_2, \sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (f_{\rho\sigma} \circ p) \right) \\
 &= \sum_{\mu,v=1}^2 p^{\mu}_1 p^v_2 \sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j \det(f_{\mu} \circ p, f_v \circ p, f_{\rho\sigma} \circ p) \\
 &= (\det p') \cdot \left[\sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (d_{\rho\sigma} \circ p) \right],
 \end{aligned}$$

and consequently

$$(7) \quad \sum_{i,j=1}^2 \tilde{d}_{ij} ds^i \otimes ds^j = (\det p') \sum_{i,j=1}^2 \left(\sum_{\rho,\sigma=1}^2 p^{\rho}_i p^{\sigma}_j (d_{\rho\sigma} \circ p) \right) ds^i \otimes ds^j.$$

If it weren't for the factor $(\det p')$, the tensors in (5) and (7) would already be the same. From (6) we easily see that

$$(8) \quad \begin{aligned} \det(\tilde{d}_{ij}) &= (\det p')^4 \cdot \det(d_{ij} \circ p) \\ \implies \sqrt[4]{\det(\tilde{d}_{ij})} &= (\det p') \cdot \sqrt[4]{\det(d_{ij} \circ p)} \quad \text{for } \det p' > 0. \end{aligned}$$

Together with (7), this gives exactly the equation (4) which we want. We have thus shown that for orientation preserving $f: U \rightarrow M$ we always have

$$\begin{aligned} \mathfrak{I}_f &= \sum_{i,j=1}^2 g_{ij} ds^i \otimes ds^j \\ \text{for } g_{ij} &= \frac{d_{ij}}{\sqrt[4]{\det(d_{ij})}}, \quad d_{ij} = \det(f_1, f_2, f_{ij}). \end{aligned}$$

This formula allows us to compare the special affine first fundamental form \mathfrak{I}_f with the ordinary *second* fundamental form \mathbf{II}_f , whose coefficients l_{ij} are

$$\begin{aligned} l_{ij} &= \langle n, f_{ij} \rangle = \left\langle \frac{f_1 \times f_2}{|f_1 \times f_2|}, f_{ij} \right\rangle \\ &= \frac{\det(f_1, f_2, f_{ij})}{|f_1 \times f_2|} = \frac{d_{ij}}{\det(g_{ij})}. \end{aligned}$$

In the classical literature, the tensor \mathfrak{I}_f is introduced a little differently. One simply notes that $\sum_{i,j} d_{ij} ds^i \otimes ds^j$ is a nice tensor to consider, because it is a special affine invariant. Then one asks whether it is also an invariant under change of parameter. After deriving equation (7) one sees that it isn't, but upon noticing (8), one realizes that dividing by $\sqrt[4]{\det(d_{ij})}$ will give a tensor that is invariant under orientation preserving change of parameter, yet still a special affine invariant.

Now consider a hyperbolic point $p \in M$, and a basis $X_1, X_2, X_3 \in \mathbb{R}^3_p$ with $X_1, X_2 \in M_p$. This again determines a matrix $S = (s_{ij})$, and we can still define an inner product on M_p by

$$\langle X_i, X_j \rangle = s_{ij},$$

but now the inner product is merely non-degenerate, and neither positive definite nor negative definite. Any other such inner product, defined for a different basis, is a constant multiple of this one. If we want these constant multiples all to be positive, then we will have to have a way of selecting a permissible direction

for the vectors X_3 . So we have to choose, arbitrarily, an orientation for M_p , and then *define* X_3 to “point inward” if and only if (X_1, X_2, X_3) is positively oriented in \mathbb{R}^3_p whenever (X_1, X_2) is positively oriented in M_p . By considering only inward pointing X_3 we obtain a class of non-degenerate inner products on M_p , any one being a positive constant multiple of any other. If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is any affine map and we give the tangent space $A(M)_{A(p)}$ the orientation which makes $A_*: M_p \rightarrow A(M)_{A(p)}$ orientation preserving, then the class of inner products on M_p is precisely A^* of the class of inner products on $A(M)_{A(p)}$.

In the present set-up it makes sense to consider ordered bases X_1, X_2 of M_p with

$$\begin{aligned}\langle X_1, X_1 \rangle &= -\langle X_2, X_2 \rangle = \text{constant} > 0 \\ \langle X_1, X_2 \rangle &= 0.\end{aligned}$$

These ordered bases will again be called “quasi-orthonormal”. For any such ordered basis, and any inward pointing $X_3 \in \mathbb{R}^3_p$, the corresponding osculating paraboloid P is given by

$$P = \{sX_1 + tX_2 + (\text{constant} > 0) \cdot (s^2 - t^2)X_3\}.$$

In terms of moving frames we have

$$S(p) = (\text{constant} > 0) \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{or} \quad \begin{cases} \psi_1^3(p) = (\text{constant} > 0) \cdot \theta^1(p) \\ \psi_2^3(p) = -(\text{constant} > 0) \cdot \theta^2(p). \end{cases}$$

From among our class of inner products on M_p we can again distinguish a particular one $\langle \cdot, \cdot \rangle_p$. We define an ordered basis X_1, X_2 to be “orthonormal”,

$$\begin{aligned}\langle X_1, X_1 \rangle_p &= -\langle X_2, X_2 \rangle_p = 1 \\ \langle X_1, X_2 \rangle_p &= 0,\end{aligned}$$

if and only if $S(p) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ for all inward pointing $X_3 \in \mathbb{R}^3_p$ satisfying $\det(X_1, X_2, X_3) = 1$. The verification that this is well-defined is similar to the case of an elliptic point. If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is a map of the form $A = T \circ B$, where T is a translation and $B: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is a linear map of determinant ± 1 , and we give the tangent space $A(M)_{A(p)}$ the orientation which makes $A_*: M_p \rightarrow A(M)_{A(p)}$ orientation preserving, then $\langle \cdot, \cdot \rangle_p$ is A^* of the inner product $\langle \cdot, \cdot \rangle_{A(p)}$ on $A(M)_{A(p)}$. Clearly an ordered basis (X_1, X_2) of M_p is orthonormal if and only if

$$\left. \begin{aligned} \psi_1^3(p) &= \theta^1(p) \\ \psi_2^3(p) &= -\theta^2(p) \end{aligned} \right\} \quad \text{or} \quad P = \left\{ sX_1 + tX_2 + \frac{1}{2}(s^2 - t^2)X_3 \right\}$$

for all inward pointing $X_3 \in \mathbb{R}^3_p$ with $\det(X_1, X_2, X_3) = 1$.

On an oriented surface M with all points hyperbolic, we now have an inner product $\langle \cdot, \cdot \rangle_p$ defined on each M_p , and hence an “indefinite Riemannian metric” $\langle \cdot, \cdot \rangle$ on M . Once again we also denote $\langle \cdot, \cdot \rangle$ by \mathfrak{I} , and call it the **special affine first fundamental form** of M . If $f: M \rightarrow \mathbb{R}^3$ is an immersion of an oriented surface with all points of the image hyperbolic, then we define the **special affine first fundamental form** \mathfrak{I}_f of f to be the tensor $\mathfrak{I}_f = f^*(\langle \cdot, \cdot \rangle)$ on M , where $\langle \cdot, \cdot \rangle$ is the special affine first fundamental form on $f(M)$, when $f(M)$ is given the orientation that makes f orientation preserving.

When $f: U \rightarrow \mathbb{R}^3$, for $U \subset \mathbb{R}^2$ open (with the usual orientation), and all points of $f(U)$ are hyperbolic, we again define functions $g_{ij}: U \rightarrow \mathbb{R}$ by

$$\mathfrak{I}_f = g_{11} ds \otimes ds + g_{12} ds \otimes dt + g_{21} dt \otimes ds + g_{22} dt \otimes dt.$$

Again take the case where $f(s, t) = (s, t, h(s, t))$, with $p = f(0, 0) = 0 \in \mathbb{R}^3$ and $M_p = (x, y)$ -plane. If $X'_j = \sum_{i=1}^2 a_{ij} X_i$ is a new basis of \mathbb{R}^2 such that P is the graph of

$$(s, t) \mapsto \frac{1}{2}(s^2 - t^2)$$

in the X'_1, X'_2, X_3 system, then we have

$$\begin{aligned} (I') \quad A^t \cdot \frac{1}{2} \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \cdot A &= \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\ \implies \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} &= (A^t)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} A^{-1} \\ \implies \det A &= (\beta^2 - \alpha\gamma)^{-1/2}. \end{aligned}$$

As before, we see that the vectors

$$X''_1 = (\sqrt[8]{\beta^2 - \alpha\gamma}) X'_1, \quad X''_2 = (\sqrt[8]{\beta^2 - \alpha\gamma}) X'_2$$

satisfy

$$\begin{aligned} \langle X''_1, X''_1 \rangle_0 &= -\langle X''_2, X''_2 \rangle_0 = 1 \\ \langle X''_1, X''_2 \rangle_0 &= 0. \end{aligned}$$

So, introducing the matrix B as before, we have

$$\begin{aligned}
 g(0,0) = \langle X_i, X_j \rangle_0 &= \frac{1}{\sqrt[4]{\beta^2 - \alpha\gamma}} \left\langle \sum_{k=1}^2 b_{ki} X''_k, \sum_{k=1}^2 b_{kj} X''_k \right\rangle_0 \\
 &= \frac{1}{\sqrt[4]{\beta^2 - \alpha\gamma}} \cdot (b_{1i}b_{1j} - b_{2i}b_{2j}) \\
 &= \frac{1}{\sqrt[4]{\beta^2 - \alpha\gamma}} \left[B^t \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} B \right]_{ij} \\
 &= \frac{1}{\sqrt[4]{\beta^2 - \alpha\gamma}} \left[(A^t)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} A^{-1} \right]_{ij}.
 \end{aligned}$$

Thus (I') gives

$$g_{ij}(0,0) = \frac{d_{ij}}{\sqrt[4]{-\det(d_{ij})}}(0,0).$$

The same calculations as before show that this formula holds for any $f: U \rightarrow \mathbb{R}^3$. We can refer to the elliptic and hyperbolic cases jointly by means of equation

$$\begin{aligned}
 \text{(I)} \quad \mathfrak{I}_f &= \sum_{i,j=1}^2 g_{ij} ds^i \otimes ds^j \\
 \text{for } g_{ij} &= \frac{d_{ij}}{\sqrt[4]{|\det(d_{ij})|}}, \quad d_{ij} = \det(f_1, f_2, f_{ij}).
 \end{aligned}$$

If M consists entirely of elliptic points, then the map $f: U \rightarrow M$ must be orientation preserving when M is given its natural orientation, and if M consists entirely of hyperbolic points, then M must be given the orientation which makes f orientation preserving. (Henceforth we will not bother to mention the subsidiary conditions on orientation which must be added to all our considerations.)

Finally, what do we do at points $p \in M$ which are flat (parabolic or planar)? The answer is, we don't do anything. We do not define $\langle \cdot, \cdot \rangle$ and *we cannot expect to*. To see that this must be so, just consider the surface $\mathbb{R}^2 \subset \mathbb{R}^3$, consisting entirely of flat points. If we could define a metric $\langle \cdot, \cdot \rangle$ on \mathbb{R}^2 which was a special affine invariant, then we would have to have $A^* \langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle$ for every special linear affine map A from \mathbb{R}^2 to \mathbb{R}^2 , since such an A can always be extended to a special linear affine map $A': \mathbb{R}^3 \rightarrow \mathbb{R}^3$. But of course, there is no metric on \mathbb{R}^2 which is invariant under all special linear affine maps. In affine geometry we will simply always assume, without explicitly mentioning it again,

that all surfaces have no flat points. Thus connected surfaces will consist either entirely of elliptic points, or entirely of hyperbolic points.

After all this work, we are hardly any closer to the problem of picking out a “special affine normal”. In order to do this we will have to consider *third* order approximations to M . We will still assume that $p = 0 \in \mathbb{R}^3$ and that M_p is the (x, y) -plane. Choose some vector $X_3 \in \mathbb{R}^3$ which is not in $\mathbb{R}^2 = (x, y)$ -plane. For every basis $\mathbf{X} = (X_1, X_2)$ of \mathbb{R}^2 , we can then consider the function $h^{\mathbf{X}}$ which describes M in the X_1, X_2, X_3 coordinate system, and we can look at the quadratic and cubic polynomials

$$\begin{aligned} & \frac{1}{2}(h^{\mathbf{X}}_{11}(0,0) \cdot s^2 + 2h^{\mathbf{X}}_{12}(0,0) \cdot st + h^{\mathbf{X}}_{22}(0,0) \cdot t^2) \\ & \frac{1}{6}(h^{\mathbf{X}}_{111}(0,0) \cdot s^3 + 3h^{\mathbf{X}}_{112}(0,0) \cdot s^2t + 3h^{\mathbf{X}}_{122}(0,0) \cdot st^2 + h^{\mathbf{X}}_{222}(0,0) \cdot t^3) \end{aligned}$$

which appear in the Taylor series for $h^{\mathbf{X}}$. As on page 37, we can also define functions $\Phi^{\mathbf{X}}, \Psi^{\mathbf{X}}: \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$\begin{aligned} \Phi^{\mathbf{X}}(sX_1 + tX_2) &= \frac{1}{2}(h^{\mathbf{X}}_{11}(0,0) \cdot s^2 + 2h^{\mathbf{X}}_{12}(0,0) \cdot st + h^{\mathbf{X}}_{22}(0,0) \cdot t^2) \\ \Psi^{\mathbf{X}}(sX_1 + tX_2) &= \frac{1}{6}(h^{\mathbf{X}}_{111}(0,0) \cdot s^3 + 3h^{\mathbf{X}}_{112}(0,0) \cdot s^2t \\ & \quad + 3h^{\mathbf{X}}_{122}(0,0) \cdot st^2 + h^{\mathbf{X}}_{222}(0,0) \cdot t^3). \end{aligned}$$

All the $\Phi^{\mathbf{X}}$ are really the same function $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}$, and all the $\Psi^{\mathbf{X}}$ are really the same function $\Psi: \mathbb{R}^2 \rightarrow \mathbb{R}$. We have already checked this for the $\Phi^{\mathbf{X}}$'s, using the relation

$$(1) \quad h^{\mathbf{X}}_{\alpha\beta}(0,0) = \sum_{j,k=1}^2 a_{j\alpha}a_{k\beta}h_{jk}(0,0);$$

the check for the $\Psi^{\mathbf{X}}$'s is similar, using the easily derived relation

$$(2) \quad h^{\mathbf{X}}_{\alpha\beta\gamma}(0,0) = \sum_{j,k,l=1}^2 a_{j\alpha}a_{k\beta}a_{l\gamma}h_{jkl}(0,0).$$

Remember that these functions Φ and Ψ *do* depend on the original choice of X_3 . We would now like to ask if there is a particular choice for the direction of X_3 which will make the functions Φ and Ψ be related to each other in some especially nice way; so the real problem here is to formulate a definite question, by deciding on a suitable criterion for declaring that Φ and Ψ are nicely related.

We now find ourselves placed in a purely algebraic situation, which we can formulate as follows. A function $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ will be called **quadratic** if for a basis X_1, X_2 of \mathbb{R}^2 we have

$$\Phi(sX_1 + tX_2) = \Phi_{11}s^2 + 2\Phi_{12}st + \Phi_{22}t^2,$$

for some numbers Φ_{jk} . More generally, for an n -dimensional vector space V , a function $\Phi: V \rightarrow \mathbb{R}$ will be called **quadratic** if for a basis X_1, \dots, X_n of V we have

$$\Phi\left(\sum_{i=1}^n s^i X_i\right) = \sum_{j,k=1}^n \Phi_{jk} s^j s^k$$

for certain numbers Φ_{jk} , which it will be convenient to assume are symmetric with respect to the indices j and k . It is easy to see that if Φ has this form for one basis, then it has this form for any other basis. Indeed, if $\bar{X}_1, \dots, \bar{X}_n$ is another basis, with $\bar{X}_j = \sum_i a_{ij} X_i$, then

$$\begin{aligned} \Phi\left(\sum_{i=1}^n s^i \bar{X}_i\right) &= \Phi\left(\sum_{i=1}^n \sum_{\rho=1}^n s^i a_{\rho i} X_\rho\right) = \Phi\left(\sum_{\rho=1}^n \left(\sum_{i=1}^n s^i a_{\rho i}\right) X_\rho\right) \\ &= \sum_{j,k=1}^n \Phi_{jk} \left(\sum_{i=1}^n s^i a_{ji}\right) \left(\sum_{i=1}^n s^i a_{ki}\right) \\ &= \sum_{\alpha,\beta=1}^n \left(\sum_{j,k=1}^n \Phi_{jk} a_{j\alpha} a_{k\beta}\right) s^\alpha s^\beta = \sum_{\alpha,\beta=1}^n \bar{\Phi}_{\alpha\beta} s^\alpha s^\beta, \end{aligned}$$

where the $\bar{\Phi}_{\alpha\beta}$ are given by

$$(3) \quad \bar{\Phi}_{\alpha\beta} = \sum_{j,k=1}^n \Phi_{jk} a_{j\alpha} a_{k\beta}.$$

Naturally, (3) is just the n -dimensional analogue of the equations (1) which gave us a well-defined Φ in the first place. Notice that in terms of the matrix $A = (a_{ij})$ and the matrices $[\Phi] = (\Phi_{jk})$ and $[\bar{\Phi}] = (\bar{\Phi}_{\alpha\beta})$ we can write (3) as

$$(4) \quad [\bar{\Phi}] = A^t \cdot [\Phi] \cdot A.$$

We will define a function $\Psi: V \rightarrow \mathbb{R}$ to be **cubic** if for a basis X_1, \dots, X_n of V we have

$$\Psi\left(\sum_{i=1}^n s^i X_i\right) = \sum_{j,k,l=1}^n \Psi_{jkl} s^j s^k s^l,$$

for certain numbers Ψ_{jkl} , which we can assume are symmetric in the indices. For the basis \bar{X} we easily find that

$$\Psi\left(\sum_{i=1}^n s^i \bar{X}_i\right) = \sum_{\alpha, \beta, \gamma=1}^n \bar{\Psi}_{\alpha\beta\gamma} s^\alpha s^\beta s^\gamma,$$

where

$$(5) \quad \bar{\Psi}_{\alpha\beta\gamma} = \sum_{j,k,l=1}^n \Psi_{jkl} a_{j\alpha} a_{k\beta} a_{l\gamma};$$

this equation, of course, is just the analogue of (2). It is not hard to see that we could also define a quadratic function $\Phi: V \rightarrow \mathbb{R}$ to be one of the form $\Phi(X) = B(X, X)$ for a symmetric bilinear function $B: V \times V \rightarrow \mathbb{R}$. Similarly, a cubic function $\Psi: V \rightarrow \mathbb{R}$ is one of the form $\Psi(X) = T(X, X, X)$ for a symmetric trilinear function $T: V \times V \times V \rightarrow \mathbb{R}$.

Now we would like to find some quantity depending on Φ and Ψ , but *not* on the choice of basis, and hence not on the particular coefficients Φ_{jk} and Ψ_{jkl} for this basis. As a warm-up, let's first take the case of two *quadratic* forms Φ and Θ . We will assume that the first quadratic form Φ is non-degenerate, by which we mean that the corresponding symmetric bilinear form is non-degenerate. More concretely, this means that if we choose a basis X_1, \dots, X_n for V , then the matrix $[\Phi] = (\Phi_{jk})$ is non-singular, and therefore has an inverse matrix $[\Phi]^{-1} = (\Phi^{jk})$ with

$$\sum_{k=1}^n \Phi_{jk} \Phi^{kl} = \delta_j^l.$$

Now consider the number

$$\sum_{j,k=1}^n \Phi^{jk} \Theta_{jk}.$$

If $\bar{X}_j = \sum_i a_{ij} X_i$ is a new basis, and $B = (b_{ij})$ is the inverse of $A = (a_{ij})$, then (4) shows that the matrix $[\bar{\Phi}] = (\bar{\Phi}_{\alpha\beta})$ satisfies

$$[\bar{\Phi}] = A^t [\Phi] A \implies [\bar{\Phi}]^{-1} = B [\Phi]^{-1} B^t$$

and hence

$$(6) \quad \bar{\Phi}^{\alpha\beta} = \sum_{j,k=1}^n \Phi^{jk} b_{\alpha j} b_{\beta k}.$$

Applying (3) to Θ , we thus find that

$$\begin{aligned}
 \sum_{\alpha, \beta=1}^n \bar{\Phi}^{\alpha\beta} \bar{\Theta}_{\alpha\beta} &= \sum_{\alpha, \beta=1}^n \sum_{j, k=1}^n \Phi^{jk} b_{\alpha j} b_{\beta k} \sum_{l, m=1}^n \Theta_{lm} a_{l\alpha} a_{m\beta} \\
 &= \sum_{j, k=1}^n \sum_{l, m=1}^n \Phi^{jk} \Theta_{lm} \delta_{lj} \delta_{mk} \\
 &= \sum_{j, k=1}^n \Phi^{jk} \Theta_{jk}.
 \end{aligned}$$

We thus have a well-defined number (Φ, Θ) , which for any basis $\{X_i\}$ is given by

$$(\Phi, \Theta) = \sum_{j, k=1}^n \Phi^{jk} \Theta_{jk}.$$

Naturally, there must also be some invariant definition of (Φ, Θ) lurking around. Indeed, the bilinear function corresponding to Φ is just what we usually call an inner product $\langle \cdot, \cdot \rangle$ on V . This gives rise to inner products on just about every other vector space in sight which is related to V . In particular, we could define an inner product $\langle \cdot, \cdot \rangle$ on the set of all bilinear functions $C: V \times V \rightarrow \mathbb{R}$; the quadratic function Θ corresponds to such a C , and (Φ, Θ) is just $\langle C, C \rangle$. But in this case I don't think it's worth all the linear algebra which this involves; it's easier to just do the calculation.

To get a little more feeling for what the number (Φ, Θ) is, take the case where the inner product corresponding to Φ is actually positive definite. Then there is a basis X_1, \dots, X_n of V such that $\Phi_{ij} = \delta_{ij}$ and also $\Theta_{ij} = 0$ for $i \neq j$ (compare Proposition II.4-14). So (Φ, Θ) is just the sum of the "diagonal terms", $(\Phi, \Theta) = \sum_i \Theta_{ii}$. When $(\Phi, \Theta) = 0$, the forms Φ and Θ are said to be **apolar**, a term that comes from the old invariant theory. We can give a very concrete geometric meaning to apolarity when V is 2-dimensional. For our special choice of basis, we see that Θ is apolar to Φ if and only if the graph of $\Theta: V \rightarrow \mathbb{R}$ is a hyperbolic paraboloid,

$$\begin{aligned}
 \Theta(sX_1 + tX_2) &= \Theta_{11}s^2 + \Theta_{22}t^2 \\
 &= \Theta_{11}s^2 - \Theta_{11}t^2,
 \end{aligned}$$

for which the set $\Theta^{-1}(0) \subset V$ is a pair of straight lines making angles of $\pi/4$ with the lines spanned by X_1 and X_2 (we measure angles in V by using the inner product on V corresponding to Φ , so that X_1 and X_2 are orthonormal).

We can also express the apolarity of Φ and Θ in a way that does not involve the special choice of basis: Φ and Θ are apolar if and only if $\Theta^{-1}(0)$ is the union of two straight lines which are perpendicular to each other in the inner product on V corresponding to Φ .

Now consider a quadratic function Φ , which we will still assume is non-degenerate, and a cubic function Ψ . Instead of constructing a number from Φ and Ψ , we will construct an element of V^* , namely

$$X_i \mapsto \sum_{j,k=1}^n \Phi^{jk} \Psi_{ijk}.$$

Suppose we have a new basis $\bar{X}_j = \sum_i a_{ij} X_i$, and that we let $B = (b_{ij})$ be the inverse matrix of $A = (a_{ij})$, as before. Consider the map

$$\bar{X}_\alpha \mapsto \sum_{\beta,\gamma=1}^n \bar{\Phi}^{\beta\gamma} \bar{\Psi}_{\alpha\beta\gamma}.$$

This map takes $X_i = \sum_\alpha b_{\alpha i} \bar{X}_\alpha$ to

$$\begin{aligned} \sum_{\alpha=1}^n b_{\alpha i} \sum_{\beta,\gamma=1}^n \bar{\Phi}^{\beta\gamma} \bar{\Psi}_{\alpha\beta\gamma} &= \sum_{\alpha=1}^n b_{\alpha i} \sum_{\beta,\gamma=1}^n \sum_{j,k=1}^n \Phi^{jk} b_{\beta j} b_{\gamma k} \sum_{l,p,q=1}^n \Psi_{lpq} a_{l\alpha} a_{p\beta} a_{q\gamma} \\ &\quad \text{by (5) and (6)} \\ &= \sum_{j,k=1}^n \Phi^{jk} \Psi_{ijk}. \end{aligned}$$

Thus we have a well-defined map $(\Phi, \Psi): V \rightarrow \mathbb{R}$, which for any basis $\{X_i\}$ is given by

$$(\Phi, \Psi)(X_i) = \sum_{j,k=1}^n \Phi^{jk} \Psi_{ijk};$$

industrious readers can supply their own invariant definition. As before, we say that Φ and Ψ are **apolar** if $(\Phi, \Psi) = 0$. Suppose that V is 2-dimensional, with a basis X_1, X_2 . Since

$$\begin{pmatrix} \Phi^{11} & \Phi^{12} \\ \Phi^{21} & \Phi^{22} \end{pmatrix} = \frac{1}{\det[\Phi]} \begin{pmatrix} \Phi_{22} & -\Phi_{21} \\ -\Phi_{12} & \Phi_{11} \end{pmatrix},$$

and the Φ_{jk} and Ψ_{jkl} are symmetric in the indices, we see that Φ and Ψ are apolar if and only if they satisfy the **apolarity conditions**

$$(*) \quad \begin{cases} \Phi_{22} \Psi_{111} - 2\Phi_{12} \Psi_{112} + \Phi_{11} \Psi_{122} = 0 \\ \Phi_{22} \Psi_{112} - 2\Phi_{12} \Psi_{122} + \Phi_{11} \Psi_{222} = 0. \end{cases}$$

Later on we will give a geometric interpretation of apolarity when V is 2-dimensional, but for the moment, we content ourselves with the observation that apolarity is clearly just about the simplest well-defined relationship that one can posit between a quadratic and a cubic function. It also happens to do the trick:

12. PROPOSITION. Let M be a surface in \mathbb{R}^3 and $p \in M$ a point which is elliptic or hyperbolic. For each tangent vector $X_3 \in \mathbb{R}^3_p$ which is not in M_p , define a quadratic function $\Phi: M_p \rightarrow \mathbb{R}$ and a cubic function $\Psi: M_p \rightarrow \mathbb{R}$ by looking at the second and third order terms in the Taylor series for the function which describes M in terms of the X_1, X_2, X_3 coordinate system, for any basis X_1, X_2 of M_p . Then there is a unique direction for X_3 which makes Φ and Ψ apolar.

PROOF. For simplicity we will assume that $p = 0 \in \mathbb{R}^3$ and that M_p is the (x, y) -plane. We can choose a basis X_1, X_2 for \mathbb{R}^2 so that if

$$(1) \quad M = \{sX_1 + tX_2 + h(s, t) \cdot (0, 0, 1)\},$$

then h has the form

$$(2) \quad h(s, t) = \frac{A}{2}(s^2 \pm t^2) + \frac{1}{6}(Bs^3 + 3Cs^2t + 3Dst^2 + Et^3) + R(s, t),$$

where $R(s, t)/|(s, t)|^3 \rightarrow 0$ as $(s, t) \rightarrow 0$.

Now consider the basis $X_1, X_2, X_3 = (0, 0, 1) + \lambda X_1 + \mu X_2$, for two constants λ, μ . Let k be the function describing M in the X_1, X_2, X_3 coordinate system, so that

$$(3) \quad \begin{aligned} M &= \{sX_1 + tX_2 + k(s, t) \cdot [(0, 0, 1) + \lambda X_1 + \mu X_2]\} \\ &= \{[s + \lambda \cdot k(s, t)]X_1 + [t + \mu \cdot k(s, t)]X_2 + k(s, t) \cdot (0, 0, 1)\}. \end{aligned}$$

Comparing with (1), we see that

$$k(s, t) = h(s + \lambda \cdot k(s, t), t + \mu \cdot k(s, t)).$$

Using (2), and noting that we have $k(s, t)/|(s, t)| \rightarrow 0$, we find that

$$\begin{aligned} k(s, t) &= \frac{A}{2}(s^2 \pm t^2) + \frac{1}{6}(Bs^3 + 3Cs^2t + 3Dst^2 + Et^3) \\ &\quad + Ak(s, t)(\lambda \cdot s \pm \mu \cdot t) + R'(s, t). \end{aligned}$$

where $R'(s, t)/|(s, t)|^3 \rightarrow 0$ as $(s, t) \rightarrow 0$. If we plug the whole right side of this equation into the term $k(s, t)$ on the right, we then find that

$$k(s, t) = \frac{A}{2}(s^2 \pm t^2) + \frac{1}{6}(Bs^3 + 3Cs^2t + 3Dst^2 + Et^3) \\ + \frac{1}{2}(s^2 \pm t^2)(\lambda \cdot s \pm \mu \cdot t) + R''(s, t),$$

where $R''(s, t)/|(s, t)|^3 \rightarrow 0$. Therefore the Φ and Ψ for the basis X_1, X_2, X_3 are given by

$$\Phi(sX_1 + tX_2) = \frac{A}{2}(s^2 \pm t^2) \\ \Psi(sX_1 + tX_2) = \frac{1}{6}([B + 3\lambda]s^3 + 3[C \pm \mu]s^2t + 3[D \pm \lambda]st^2 + [E + 3\mu]t^3).$$

The apolarity conditions $(*)$ for this Φ and Ψ are

$$\pm(B + 3\lambda) + (D \pm \lambda) = 0 \\ \pm(C \pm \mu) + (E + 3\mu) = 0.$$

There are clearly unique λ and μ for which these equations hold. It is also clear that if the apolarity conditions hold for some X_3 , then they also hold for any multiple of X_3 . So there is a unique direction for X_3 which will make Φ and Ψ apolar. \diamond

For an elliptic or hyperbolic point $p \in M$, the unique direction through p which is given by Proposition 12 is called the **affine normal direction** at p . It is clearly a general affine invariant: If $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is any affine map, then the affine normal direction through the point $A(p) \in A(M)$ is the image under A of the affine normal direction through $p \in M$.

The affine normal direction can also be characterized in terms of moving frames. For simplicity we first assume that $p = 0 \in \mathbb{R}^3$ and that $M_p = (x, y)$ -plane. If the vector v points along the affine normal direction through p , then M is the graph, in the $(1, 0, 0), (0, 1, 0), v$ coordinate system, of a function h satisfying the apolarity conditions $(*)$:

$$0 = h_{22}h_{111} - 2h_{12}h_{112} + h_{11}h_{122} = \frac{\partial}{\partial s} \det h_{ij} \\ 0 = h_{22}h_{112} - 2h_{12}h_{122} + h_{11}h_{222} = \frac{\partial}{\partial t} \det h_{ij} \quad \text{at } (0, 0).$$

We can write these two equations together simply as

$$(1) \quad 0 = d \det(h_{ij}) \quad \text{at } (0, 0).$$

Locally M is the image of the map $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by

$$f(s, t) = (s, t, 0) + h(s, t) \cdot v.$$

First take the particularly simple moving frame

$$\overset{\circ}{X}_1 = f_1 = (1, 0, 0) + h_1 v, \quad \overset{\circ}{X}_2 = f_2 = (0, 1, 0) + h_2 v, \quad \overset{\circ}{X}_3 = v$$

and let $\overset{\circ}{\psi}_\beta^\alpha$ be the corresponding connection forms. Then

$$\begin{aligned} 0 \overset{\circ}{X}_1 + 0 \overset{\circ}{X}_2 + h_{ik} \overset{\circ}{X}_3 &= \nabla'_{\overset{\circ}{X}_k} \overset{\circ}{X}_i = \sum_{\alpha=1}^3 \overset{\circ}{\psi}_i^\alpha(\overset{\circ}{X}_k) \cdot \overset{\circ}{X}_\alpha \quad i, k = 1, 2 \\ 0 &= \nabla'_{\overset{\circ}{X}_k} \overset{\circ}{X}_3 = \sum_{\alpha=1}^3 \overset{\circ}{\psi}_3^\alpha(\overset{\circ}{X}_k) \cdot \overset{\circ}{X}_\alpha \quad k = 1, 2, \end{aligned}$$

which implies, among other things, that

$$(2) \quad \begin{aligned} \overset{\circ}{\psi}_i^j &= 0 \quad \text{on } TM \quad i, j = 1, 2 \\ \overset{\circ}{\psi}_3^\alpha &= 0 \quad \text{on } TM \quad \alpha = 1, 2, 3. \end{aligned}$$

Now consider an arbitrary adapted moving frame X_1, X_2, X_3 subject only to the condition that

$$X_3 = \overset{\circ}{X}_3 \quad \text{at } p.$$

This condition implies that the matrix a on page 80 relating the two frames also satisfies

$$(3) \quad \begin{aligned} a_{i3}(p) &= 0 \quad i = 1, 2; & a_{33}(p) &= 1 \\ \implies (a^{-1})_{i3}(p) &= 0 \quad i = 1, 2; & (a^{-1})_{33}(p) &= 1. \end{aligned}$$

From the general equation (d) on page 81 we have

$$\begin{aligned} \psi_i^j &= (a^{-1} da)_i^j + (a^{-1} \overset{\circ}{\psi} a)_i^j \\ &= \sum_{\alpha=1}^3 (a^{-1})_{j\alpha} da_{\alpha i} + \sum_{\alpha, \beta=1}^3 (a^{-1})_{j\alpha} \overset{\circ}{\psi}_\beta^\alpha a_{\beta i} \quad i, j = 1, 2 \\ \psi_3^3 &= (a^{-1} da)_3^3 + (a^{-1} \overset{\circ}{\psi} a)_3^3 \\ &= \sum_{\alpha=1}^3 (a^{-1})_{3\alpha} da_{\alpha 3} + \sum_{\alpha, \beta=1}^3 (a^{-1})_{3\alpha} \overset{\circ}{\psi}_\beta^\alpha a_{\beta 3}. \end{aligned}$$

Taking into account equations (2) and (3), as well as equation (b) on page 80, we obtain

$$(4) \quad \psi_3^3 = da_{33} \quad \text{on } TM, \text{ at } p$$

and

$$\psi_i^j = \sum_{k=1}^2 (a^{-1})_{jk} da_{ki} \quad \text{on } TM, \text{ at } p \quad i, j = 1, 2;$$

since our matrix a has the form

$$\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{at } p,$$

we can write the latter equation as

$$(5) \quad \psi_i^j = \sum_{k=1}^2 (A^{-1})_{jk} dA_{ki} \quad \text{on } TM, \text{ at } p \quad i, j = 1, 2,$$

where A is the 2×2 matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. But we can compute da_{33} in terms of the matrix S for the moving frame X_1, X_2, X_3 . For equation (e) on page 81 relates S to $\overset{\circ}{S} = (h_{ij})$ by

$$\begin{aligned} a_{33} \cdot S &= A^t \overset{\circ}{S} A \implies (a_{33})^2 \cdot (\det S) = (\det A)^2 \cdot (\det \overset{\circ}{S}) \\ &\implies 2 \log |a_{33}| + \log |\det S| = 2 \log |\det A| + \log |\det \overset{\circ}{S}|. \end{aligned}$$

Since $a_{33}(p) = 1$, at p we have

$$\begin{aligned} 2 da_{33} + d \log |\det S| &= 2d \log |\det A| + d \log |\det \overset{\circ}{S}| \\ &= 2d \log |\det A|, \quad \text{by (I)}. \end{aligned}$$

Hence (4) becomes

$$(6) \quad \psi_3^3 = d \log |\det A| - \frac{1}{2} d \log |\det S| \quad \text{on } TM, \text{ at } p.$$

Now note that

$$\begin{aligned} d \log |\det A| &= \frac{d \det A}{\det A} = \frac{d(A_{11}A_{22} - A_{12}A_{21})}{\det A} \\ &= \frac{(A_{22} dA_{11} - A_{12} dA_{21}) + (-A_{21} dA_{12} + A_{11} dA_{22})}{\det A}. \end{aligned}$$

Since

$$A^{-1} = \frac{\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}}{\det A},$$

this becomes

$$\begin{aligned} d \log |\det A| &= \sum_{k=1}^2 (A^{-1})_{1k} dA_{k1} + \sum_{k=1}^2 (A^{-1})_{2k} dA_{k2} \\ &= \psi_1^1 + \psi_2^2 \quad \text{on } TM, \text{ at } p \quad \text{by (5).} \end{aligned}$$

Substituting into (6) we see that

An adapted moving frame X_1, X_2, X_3 on $M \subset \mathbb{R}^3$ has $X_3(p)$ pointing in the affine normal direction at p if and only if

$$\psi_3^3 = \psi_1^1 + \psi_2^2 - \frac{1}{2} d \log |\det S| \quad \text{on } TM, \text{ at } p.$$

From this fact we see, what is by no means *a priori* clear, that the condition $\psi_3^3 = \psi_1^1 + \psi_2^2 - \frac{1}{2} d \log |\det S|$ on TM at p depends only on the direction of X_3 at p . It is also possible to check this by a direct, quite unpleasant, calculation. If we had somehow independently observed this fact, we could have used this equation to *define* the affine normal direction. Of course, it is hard to see how one would ever be led to such an “observation”, but there is at least a way to simplify the equation. We have already observed that the condition

$$S(p) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{or} \quad \begin{cases} \psi_1^3 = \theta^1 \\ \psi_2^3 = \pm \theta^2 \end{cases}$$

depends only on the value of the moving frame X_1, X_2, X_3 at p , so it is reasonable to restrict our attention to frames satisfying this condition at all points. It is clear that

An adapted moving frame X_1, X_2, X_3 on $M \subset \mathbb{R}^3$ with $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ everywhere, has $X_3(p)$ pointing in the affine normal direction at p if and only if

$$\psi_3^3 = \psi_1^1 + \psi_2^2 \quad \text{on } TM, \text{ at } p.$$

The direct verification that for such frames the condition $\psi_3^3 = \psi_1^1 + \psi_2^2$ on TM at p depends only on the direction of X_3 at p is a little easier, and gives a nice

criterion for the direction of the affine normal; of course, as a definition it still leaves a lot to be desired in terms of motivation.

Now that we have picked out an affine normal direction which is a general affine invariant, it is a simple matter to define a special affine normal vector which is a special affine invariant. We define the **special affine normal** ν_p of M at p to be the unique vector $\nu_p \in \mathbb{R}^3_p$ such that

- (1) ν_p points along the affine normal direction at p
- (2) $\det(X_1, X_2, \nu_p) = 1$ for every positively oriented basis X_1, X_2 of M_p which is orthonormal for the metric $\langle \cdot, \cdot \rangle_p$.

If $f: M \rightarrow \mathbb{R}^3$ is an immersion, we let \mathcal{N} be the vector field along f such that $\mathcal{N}(p)_{f(p)} \in \mathbb{R}^3_{f(p)}$ is the special affine normal to $f(M)$ at $f(p)$. We will not try to derive a formula for \mathcal{N} right away, since it will come out very naturally later on.

Suppose now that we have an adapted moving frame X_1, X_2, X_3 on M such that $\det(X_1, X_2, X_3) = 1$. Then for all tangent vectors X to M we have

$$\begin{aligned} 0 &= X(\det(X_1, X_2, X_3)) \\ &= \det(\nabla'_X X_1, X_2, X_3) + \det(X_1, \nabla'_X X_2, X_3) + \det(X_1, X_2, \nabla'_X X_3) \\ &= \det\left(\sum_{\alpha=1}^3 \psi_1^\alpha(X) \cdot X_\alpha, X_2, X_3\right) + \cdots \\ &= \psi_1^1(X) + \psi_2^2(X) + \psi_3^3(X). \end{aligned}$$

Thus we have

$$\psi_1^1 + \psi_2^2 + \psi_3^3 = 0 \quad \text{on } TM.$$

Suppose, moreover, that (X_1, X_2) is positively oriented and orthonormal with respect to the metric $\langle \cdot, \cdot \rangle$ on M ; in other words, suppose that $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ everywhere. We have seen that in this case $X_3(p)$ points in the affine normal direction if and only if

$$\psi_3^3 = \psi_1^1 + \psi_2^2 \quad \text{on } TM, \text{ at } p.$$

Since we also have $\psi_1^1 + \psi_2^2 + \psi_3^3 = 0$ everywhere on TM , we find that

An adapted moving frame X_1, X_2, X_3 on $M \subset \mathbb{R}^3$ with (X_1, X_2) positively oriented, $\det(X_1, X_2, X_3) = 1$, and $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ everywhere, has $X_3(p) = \nu_p$ if and only if

$$\psi_3^3 = 0 \quad \text{on } TM, \text{ at } p.$$

Now the condition $\psi_3^3(p) = 0$ on M_p means that

$$\nabla'_{X_p} X_3 \in M_p \quad \text{for all } X_p \in M_p.$$

So we see, in particular, that we always have

$$\nabla'_{X_p} \nu \in M_p \quad \text{for } X_p \in M_p,$$

just as we always have $\nabla'_{X_p} \nu \in M_p$ in ordinary surface theory. (Verifying this fact directly from the definition of ν involves a rather hideous computation, and in general all the formulas and computations for affine surface theory are considerably more complicated than those in ordinary surface theory; that is why we have brought in moving frames, with their attendant computational simplifications, so early in the game.) We will denote the map $X_p \mapsto \nabla'_{X_p} \nu$ from M_p to M_p by $d\nu: M_p \rightarrow M_p$. The reason for this notation is the same as in ordinary surface theory: since we have a vector $\nu_p \in \mathbb{R}^3_p$ for each $p \in M$, we have a map $\nu: M \rightarrow \mathbb{R}^3$, and $\nabla'_{X_p} \nu$ is the same as $[d\nu(X_p)]_p$. In the present case, the map ν doesn't go into any special subset of \mathbb{R}^3 , but $\nu(M) \subset \mathbb{R}^3$ will be some surface, at least near the points where $d\nu: M_p \rightarrow M_p$ is non-singular. The relation $\nabla'_{X_p} \nu \in M_p$ tells us that the tangent plane M_p is parallel to the tangent plane $\nu(M)_{\nu(p)}$. So we could also denote $d\nu: M_p \rightarrow M_p$ by $\nu_*: M_p \rightarrow M_p$, as in the case of ordinary surface theory.

Before proceeding with the development of special affine surface theory, we pause briefly to describe the procedure usually found in those papers and books which present the theory totally from the moving frame point of view, and ignore the question of general affine invariants, like the affine normal direction. One works, first of all, only with adapted frames X_1, X_2, X_3 satisfying $\det(X_1, X_2, X_3) = 1$, and hence $\psi_1^1 + \psi_2^2 + \psi_3^3 = 0$ on TM . Equation (e) is first derived, except that now a_{33} must = 1; it shows that S is determined only up to a transformation $S \mapsto A^t S A$. The canonical forms for symmetric matrices under this equivalence relation are precisely $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$; so one "normalizes" the frame by requiring that S be everywhere of this form. Then one further normalizes the frame by requiring that $\psi_3^3(p) = 0$ for all p , noting that the condition $\psi_3^3(p) = 0$ now uniquely determines $X_3(p)$; indeed if $\{X_\alpha\}$ and $\{X'_\alpha\}$ are two adapted frames with $\det(X_1, X_2, X_3) = \det(X'_1, X'_2, X'_3) = 1$ and $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ everywhere, so that

$$\begin{aligned} \psi_1^3 &= \theta^1 \\ \psi_2^3 &= \pm \theta^2, \end{aligned}$$

then the matrix a relating the X'_α to the X_α must be of the form

$$a = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

and equation (d) on page 81 yields

$$\begin{aligned} \psi'_3 &= \psi_1^3 a_{13} + \psi_2^3 a_{23} + \psi_3^3 \\ &= a_{13} \theta^1 + a_{23} \theta^2 + \psi_3^3, \end{aligned}$$

so that $\psi_3^3(p) = \psi'_3(p) = 0 \implies a_{13}(p) = a_{23}(p) = 0 \implies X'_3(p) = X_3(p)$. The uniquely determined X_3 is now dubbed the special affine normal, and any basis X_1, X_2 with $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ is called orthonormal, thereby defining the special affine first fundamental form. These normalizations are usually carried out with hardly a word of motivation, as if they are so natural that any idiot would immediately think of doing them—in reality, of course, the authors already knew what results they wanted, since they were simply reformulating a classical theory.

Now that we have finally determined the special affine normal v_p at $p \in M$, we can imitate a basic construction from ordinary surface theory. First we introduce the unique projections

$$\begin{aligned} \mathsf{T}: \mathbb{R}^3_p &\rightarrow M_p \\ \mathsf{L}: \mathbb{R}^3_p &\rightarrow \mathbb{R} \cdot v_p = \text{all multiples of } v_p \end{aligned}$$

such that

$$X = \mathsf{T}X + \mathsf{L}X \quad \text{for all } X \in M_p.$$

Notice that $\mathsf{T}: \mathbb{R}^3_p \rightarrow M_p$ is generally different from the tangential projection $\mathsf{T}: \mathbb{R}^3_p \rightarrow M$ of ordinary surface theory. Given vector fields X, Y tangent along M , we would like to find $\mathsf{T}\nabla'_X Y$ and $\mathsf{L}\nabla'_X Y$, where $\nabla'_X Y$ is the ordinary covariant differentiation in \mathbb{R}^3 . Now in ordinary surface theory, $\mathsf{L}\nabla'_X Y$ involved the second fundamental form II . Since II is so closely related to the affine first fundamental form I , we might expect I to be involved in $\mathsf{L}\nabla'_X Y$. As a matter of fact,

13. PROPOSITION. If X, Y are tangent along $M \subset \mathbb{R}^3$, then

$$\mathsf{L}\nabla'_X Y = \langle X, Y \rangle \cdot v = \mathsf{I}(X, Y) \cdot v.$$

PROOF. Note first that, as in the case of $\mathcal{L}\nabla'_X Y$, we have

$$\mathcal{L}\nabla'_{X_p} fY = \mathcal{L}(X_p(f) \cdot Y_p + f(p) \cdot \nabla'_{X_p} Y) = f(p) \cdot \mathcal{L}\nabla'_{X_p} Y,$$

so $\mathcal{L}\nabla'_{X_p} Y$ actually depends only on the value of Y at p . Now let (X_1, X_2) be a positively oriented moving frame on M which is orthonormal for $\langle \cdot, \cdot \rangle$; this means (taking the elliptic case for simplicity) that for the moving frame $(X_1, X_2, X_3) = (X_1, X_2, \nu)$ we have

$$\psi_i^3 = \theta^i \quad \text{on } TM.$$

But then

$$\mathcal{L}\nabla'_{X_p} X_i = \psi_i^3(X) \cdot \nu = \theta^i(X) \cdot \nu = \langle X, X_i \rangle \cdot \nu. \quad \blacklozenge$$

Now let's consider $\mathcal{T}\nabla'_X Y$. In ordinary surface theory this is $\nabla_X Y$, where ∇ is the connection on M determined by the metric $i^*\langle \cdot, \cdot \rangle$. In affine surface theory, we can consider the connection ∇ on M determined by the metric $\langle \cdot, \cdot \rangle$ (this exists even if $\langle \cdot, \cdot \rangle$ is not positive definite; compare pg. II.342). Now one can ask whether $\mathcal{T}\nabla'_X Y = \nabla_X Y$ for vector fields X and Y tangent along M . This simply isn't true, and we therefore define a map $\mathfrak{s}: M_p \times M_p \rightarrow M_p$ by

$$\mathcal{T}\nabla'_{X_p} Y = \nabla_{X_p} Y + \mathfrak{s}(X_p, Y_p),$$

where Y is a vector field tangent along M which extends Y_p . The notation $\mathfrak{s}(X_p, Y_p)$ is justified, since

$$\begin{aligned} \mathcal{T}\nabla'_{X_p} fY - \nabla_{X_p} fY &= \mathcal{T}(X_p(f) \cdot Y_p + f(p) \cdot \nabla'_{X_p} Y) \\ &\quad - [X_p(f) \cdot Y_p + f(p) \cdot \nabla_{X_p} Y] \\ &= f(p) \cdot [\mathcal{T}\nabla'_{X_p} Y - \nabla_{X_p} Y]. \end{aligned}$$

14. PROPOSITION. The tensor \mathfrak{s} is symmetric.

PROOF. If X and Y are extensions of $X_p, Y_p \in M_p$ to vector fields which are tangent to M at all points of M , then

$$\begin{aligned} \mathfrak{s}(X_p, Y_p) - \mathfrak{s}(Y_p, X_p) &= \mathcal{T}(\nabla'_{X_p} Y - \nabla'_{Y_p} X) - (\nabla_{X_p} Y - \nabla_{Y_p} X) \\ &= [X, Y](p) - [X, Y](p) = 0. \quad \blacklozenge \end{aligned}$$

We can now write the decomposition

$$\nabla'_X Y = \Upsilon \nabla'_X Y + \mathcal{L} \nabla'_X Y$$

in the following form:

The Special Affine Gauss Formulas:
 $\nabla'_X Y = \nabla_X Y + \mathcal{L}(X, Y) + \langle X, Y \rangle \nu$
 for vector fields X, Y tangent along M .

In ordinary surface theory we often found that the second fundamental form $\mathbb{I}(X, Y) = \langle s(X, Y), \nu \rangle$ was easier to work with than the map $s: M_p \times M_p \rightarrow M_p^\perp$ itself. Since \mathcal{L} now goes into M_p , we have to adopt a slightly different strategy. We define the **special affine second fundamental form** \mathbb{I} of M to be the *tri*-linear map

$$\mathbb{I}(X, Y, Z) = \langle \mathcal{L}(X, Y), Z \rangle.$$

Naturally, when $f: M \rightarrow \mathbb{R}^3$ is an immersion, we define the **special affine second fundamental form** \mathbb{I}_f of f to be $f^* \mathbb{I}$, where \mathbb{I} is the special affine second fundamental form of $f(M)$. For $f: U \rightarrow \mathbb{R}^3$ (with $U \subset \mathbb{R}^2$ open), we write the form \mathbb{I}_f as

$$\mathbb{I}_f = \sum_{i,j,k=1}^2 \ell_{ijk} ds^i \otimes ds^j \otimes ds^k.$$

Since the components of \mathbb{I}_f have three indices, there is no possibility of confusing the ℓ_{ijk} with the l_{ij} (if things hadn't worked out like this, I think I would have committed suicide at this point). We also write

$$\mathcal{L}(f_i, f_j) = \sum_{k=1}^2 \ell_{ij}^k f_k,$$

so that

$$\begin{aligned} \ell_{ijk} &= \langle \mathcal{L}(f_i, f_j), f_k \rangle = \left\langle \sum_{\rho=1}^2 \ell_{ij}^\rho f_\rho, f_k \right\rangle \\ &= \sum_{\rho=1}^2 g_{k\rho} \ell_{ij}^\rho; \end{aligned}$$

equivalently,

$$\ell_{ij}^k = \sum_{\rho=1}^2 g^{k\rho} \ell_{ij\rho}.$$

Suppose that (X_1, X_2) is a positively oriented moving frame on M which is orthonormal for $\langle \cdot, \cdot \rangle$, and let $X_3 = \nu$. For the moment consider the case where all points of M are elliptic. We set

$$c_{ijk} = \mathfrak{T}(X_i, X_j, X_k) = \langle \mathfrak{A}(X_i, X_j), X_k \rangle,$$

so that we have

$$\begin{aligned} \mathfrak{A}(X, Y) &= \sum_{k=1}^2 c_{ijk} \theta^i(X) \theta^j(Y) \cdot X_k \\ c_{ijk} &= c_{jik}. \end{aligned}$$

We also let ω_i^j be the connection forms for the frame X_1, X_2 which are determined by the metric $\langle \cdot, \cdot \rangle$; thus ω_i^j are the unique 1-forms on M satisfying

$$\begin{aligned} d\theta^i &= - \sum_{j=1}^2 \omega_j^i \wedge \theta^j \\ \omega_j^i &= -\omega_i^j. \end{aligned}$$

Since

$$\nabla' X_i = \nabla_X X_i + \mathfrak{A}(X_i, X),$$

we have

$$\psi_i^k(X) = \omega_i^k(X) + \sum_{j=1}^2 c_{ijk} \theta^j(X)$$

and hence

$$(1) \quad \psi_i^k = \omega_i^k + \sum_{j=1}^2 c_{ijk} \theta^j.$$

Since $\omega_i^k = -\omega_k^i$, we obtain

$$(2) \quad \psi_i^k + \psi_k^i = \sum_{j=1}^2 (c_{ijk} + c_{kji}) \theta^j.$$

(As usual, these formulas are understood as formulas on TM). From this we immediately deduce

15. PROPOSITION. The tensor \mathfrak{s} satisfies the “apolarity condition”

$$\text{trace}(X \mapsto \mathfrak{s}(X, Y)) = 0.$$

In terms of a map $f: U \rightarrow \mathbb{R}^3$ we have

$$0 = \sum_{i,k=1}^2 g^{ik} \ell_{ijk} = \sum_{i,k=1}^2 g^{ik} \ell_{jik};$$

in other words, the quadratic form determined by $\langle \cdot, \cdot \rangle_p$ is apolar to the cubic form determined by $\mathfrak{I}(p)$.

PROOF. For our moving frame (X_1, X_2, X_3) we have $\psi_1^1 + \psi_2^2 + \psi_3^3 = 0$, and also $\psi_3^3 = 0$, so $\psi_1^1 + \psi_2^2 = 0$. Using equation (2), this gives

$$\begin{aligned} 0 = \sum_{j=1}^2 \left(\sum_{i=1}^2 c_{iji} \right) \theta^j &\implies 0 = \sum_{i=1}^2 c_{iji} \\ &= \sum_{i=1}^2 \langle \mathfrak{s}(X_i, X_j), X_i \rangle \\ &= \text{trace}(X \mapsto \mathfrak{s}(X, X_j)). \end{aligned}$$

Similar, but slightly more involved, computations give the same result in the hyperbolic case. The second part of the Proposition is merely a restatement of the first, as we easily compute by using Fact 0. ♦

By bringing in the structural equations, we obtain one other important piece of information.

16. PROPOSITION. The tensor \mathfrak{I} is symmetric in all three arguments; equivalently,

$$\langle \mathfrak{s}(X, Y), Z \rangle = \langle \mathfrak{s}(X, Z), Y \rangle.$$

In terms of a map $f: U \rightarrow \mathbb{R}^3$ we have

$$\ell_{ijk} = \ell_{ikj}.$$

PROOF. Again we consider only the elliptic case, and leave the hyperbolic case to the reader. From equation (2) on page 106 we obtain

$$(1) \quad \sum_{k=1}^2 \psi_i^k \wedge \theta^k + \sum_{k=1}^2 \psi_k^i \wedge \theta^k = \sum_{j,k=1}^2 (c_{ijk} + c_{kji}) \theta^j \wedge \theta^k.$$

Now

$$(2) \quad \sum_{k=1}^2 \psi_k^i \wedge \theta^k = -d\theta^i$$

by the first structural equation. But we also have $\psi_i^3 = \theta^i$, so we also get

$$(3) \quad \begin{aligned} d\theta^i &= d\psi_i^3 = -\sum_{k=1}^2 \psi_k^3 \wedge \psi_i^k \quad (\text{since } \psi_3^3 = 0) \\ &= -\sum_{k=1}^2 \theta^k \wedge \psi_i^k = \sum_{k=1}^2 \psi_i^k \wedge \theta^k. \end{aligned}$$

From (1), (2), (3) we have

$$0 = \sum_{j,k=1}^2 (c_{ijk} + c_{kji}) \theta^j \wedge \theta^k,$$

so for each i, j, k we have

$$\begin{aligned} 0 &= c_{ijk} + c_{kji} - c_{ikj} - c_{jki} \\ &= c_{ijk} - c_{ikj}. \quad \spadesuit \end{aligned}$$

From the apolarity conditions we can now easily obtain an explicit formula for the vector field \mathcal{N} along a map $f: U \rightarrow \mathbb{R}^3$. We write the special affine Gauss formulas as

$$\begin{aligned} f_{ij} &= \nabla_{f_i} f_j + \mathfrak{s}(f_i, f_j) + g_{ij} \mathcal{N} \\ &= \sum_{k=1}^2 \Gamma_{ij}^k f_k + \sum_{k=1}^2 \ell_{ij}^k f_k + g_{ij} \mathcal{N}, \end{aligned}$$

where the Γ_{ij}^k are the Christoffel symbols for the metric $\langle \cdot, \cdot \rangle$, and hence computable in terms of the g_{ij} . Now

$$\sum_{i,j=1}^2 g^{ij} g_{ij} = 2,$$

so we have

$$\begin{aligned}\mathcal{N} &= \frac{1}{2} \sum_{i,j=1}^2 g^{ij} \left(f_{ij} - \sum_{k=1}^2 \Gamma_{ij}^k f_k - \sum_{k=1}^2 \ell_{ij}^k f_k \right) \\ &= \frac{1}{2} \sum_{i,j=1}^2 g^{ij} \left(f_{ij} - \sum_{k=1}^2 \Gamma_{ij}^k f_k \right) - \frac{1}{2} \sum_{i,j=1}^2 \sum_{k,\rho=1}^2 g^{ij} g^{k\rho} \ell_{ij\rho} \\ &= \frac{1}{2} \sum_{i,j=1}^2 g^{ij} \left(f_{ij} - \sum_{k=1}^2 \Gamma_{ij}^k f_k \right),\end{aligned}$$

using Proposition 15 and the fact that $\ell_{ij\rho} = \ell_{\rho ij}$, by Proposition 16. This equation means that each component \mathcal{N}^α of \mathcal{N} is given by

$$(II) \quad \mathcal{N}^\alpha = \frac{1}{2} \sum_{i,j=1}^2 g^{ij} \left(f_{ij}^\alpha - \sum_{k=1}^2 \Gamma_{ij}^k f_k^\alpha \right) \quad \alpha = 1, 2, 3.$$

Notice that the partial derivatives f_i^α are the components of the vector field df^α on U . So the tensor $\nabla(df^\alpha)$ of type $\binom{2}{0}$ has components (pg. II.210)

$$f_{i;j}^\alpha = f_{ij}^\alpha - \sum_{k=1}^2 \Gamma_{ij}^k f_k^\alpha.$$

[In Chapter 1 we wrote $f_{;i}$ instead of f_i for $\partial f / \partial x^i$, but when we are dealing with the standard coordinate system on \mathbb{R}^2 we will revert to the standard subscript notation for partial derivatives; we use $;$ rather than $'$ to emphasize that we are using the covariant derivative ∇ .] We can therefore also write

$$(II') \quad \mathcal{N}^\alpha = \frac{1}{2} \sum_{i,j=1}^2 g^{ij} f_{i;j}^\alpha.$$

Our formula is rather complicated, but it involves only quantities computable in terms of the g_{ij} 's, and hence ultimately in terms of f : in Addendum 1 to Chapter 7 we will have a little more to say about it.

With our present proof of Proposition 15, the fact that our fundamental forms \mathbf{I}, \mathbf{II} satisfy the same apolarity conditions which we used to *define* \mathbf{II} works out as some sort of miracle. So it will perhaps be reassuring to see this fact demonstrated by a computation, which will also be useful later. We will assume that $p = 0 \in \mathbb{R}^3$, the tangent plane M_p is the (x, y) -plane, and the z -axis is the affine

normal direction at p . Then M is the image of an immersion $f(s, t) = (s, t, h(s, t))$ with $h(0, 0) = h_1(0, 0) = h_2(0, 0) = 0$, and h satisfies the apolarity conditions

$$(1) \quad 0 = d(\det h_{ij}) \quad \text{at } (0, 0).$$

Now $\nabla'_{f_i} f_j = (0, 0, h_{ij})$, and the tangential projection \mathfrak{T} at 0 is just the ordinary projection on the (x, y) -plane, so $\mathfrak{T}\nabla'_{f_i} f_j = 0$ at p . Consequently, $\mathfrak{A}(f_i, f_j) = -\nabla_{f_i} f_j$ at $(0, 0)$, which gives

$$\begin{aligned} (2) \quad \ell_{ijk} &= \langle \mathfrak{A}(f_i, f_j), f_k \rangle = -\langle \nabla_{f_i} f_j, f_k \rangle && \text{at } (0, 0) \\ &= -\sum_{\rho=1}^2 \Gamma_{ij}^{\rho} \langle f_{\rho}, f_k \rangle = -\sum_{\rho=1}^2 g_{\rho k} \Gamma_{ij}^{\rho} && \text{at } (0, 0) \\ &= -[ij, k] && \text{at } (0, 0) \end{aligned}$$

where $[ij, k]$ are Christoffel symbols for \mathfrak{I}_f . To compute them, we note that

$$\begin{aligned} \frac{\partial g_{ij}}{\partial s^k} &= \frac{\partial}{\partial s^k} [d_{ij} \cdot (\det(d_{ij}))^{-1/4}] && \text{at } (0, 0) \\ &= (\det(d_{ij}))^{-1/4} \frac{\partial d_{ij}}{\partial s^k} && \text{at } (0, 0) \text{ by (1)} \\ &= (\det(d_{ij}))^{-1/4} \frac{\partial}{\partial s^k} \det(f_1, f_2, f_{ij}) && \text{at } (0, 0) \\ &= (\det(d_{ij}))^{-1/4} [\det(f_{1k}, f_2, f_{ij}) + \det(f_1, f_{2k}, f_{ij}) \\ &\quad + \det(f_1, f_2, f_{ijk})] && \text{at } (0, 0) \\ &= (\det(d_{ij}(0, 0)))^{-1/4} [0 + 0 + h_{ijk}(0, 0)], \end{aligned}$$

since each f_{ij} has its first 2 components equal 0 at $(0, 0)$. Thus

$$\begin{aligned} [ij, k](0, 0) &= \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial s^j} + \frac{\partial g_{jk}}{\partial s^i} - \frac{\partial g_{ij}}{\partial s^k} \right) (0, 0) \\ &= \frac{(h_{ikj} + h_{jki} - h_{ijk})}{2 \cdot \sqrt[4]{\det(d_{ij})}} (0, 0), \end{aligned}$$

and (2) becomes

$$(3) \quad \ell_{ijk}(0, 0) = -\frac{h_{ijk}(0, 0)}{2 \cdot \sqrt[4]{\det(d_{ij}(0, 0))}}.$$

This demonstrates Propositions 15 and 16 (since we can always make the z -axis the direction of the affine normal direction by a suitable special affine linear map, and the assertions in question are invariant under such maps).

In ordinary surface theory we used $\mathbf{I}(p)$ and $\mathbf{II}(p)$ to define two numerical invariants, the principal curvatures $k_1(p), k_2(p)$ [or equivalently $K(p)$ and $H(p)$]. These invariants arise quite naturally from an algebraic point of view, as the invariants of a 2×2 symmetric matrix S under the map $S \mapsto A^{-1}SA$, for 2×2 orthogonal matrices A . Geometrically, these invariants arise when we describe the osculating paraboloid of M (using the ordinary normal ν_p as the third axis)—for $p \in M \subset \mathbb{R}^3$ and $p' \in M' \subset \mathbb{R}^3$ we have $\{k_1(p), k_2(p)\} = \{k_1(p'), k_2(p')\}$ if and only if there is a special orthogonal affine map $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ which takes the osculating paraboloid P at p onto the osculating paraboloid P' at p' . In special affine surface theory we now want to use $\mathbf{I}(p)$ and $\mathbf{II}(p)$ to determine numerical invariants. Algebraically, we are asking for invariants of a cubic form Ψ under the map $\Psi \mapsto \bar{\Psi}$, where

$$\bar{\Psi}_{\alpha\beta\gamma} = \sum_{j,k,l=1}^2 \Psi_{jkl} a_{j\alpha} a_{k\beta} a_{l\gamma}$$

for 2×2 matrices A which preserve a certain quadratic form Φ . In the olden days when mighty invariant theory held sway over most of the domains of mathematics, this question was as natural to consider as the first, and one could probably preface the answer with the standard refrain “As every undergraduate knows . . .”. Nowadays, of course, not even graduate students know, or care, what the invariants for cubic forms are. Instead of describing them algebraically, we pass immediately to the (equivalent) geometric problem. For each $p \in M$ we have the function $\Phi + \Psi: M_p \rightarrow \mathbb{R}$ obtained by looking at the second and third order terms in the Taylor series for the function which describes M in the X_1, X_2, ν_p coordinate system, for any basis X_1, X_2 of M_p (which one doesn't matter). We will call the graph S of $\Phi + \Psi$ in the X_1, X_2, ν_p coordinate system the **osculating cubic** at p , and we want to classify these osculating cubics up to special linear affine maps.

17. PROPOSITION. Let S be the osculating cubic at a point p of a surface $M \subset \mathbb{R}^3$. If p is an elliptic point, then there is a special linear affine map $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that $A(S) \subset \mathbb{R}^3$ is the graph (in the ordinary coordinate system for \mathbb{R}^3) of the function

$$(a) \quad (s, t) \mapsto \frac{1}{2}(s^2 + t^2) + \frac{C}{6}(s^3 - 3st^2), \quad \text{for some } C.$$

If p is a hyperbolic point, then we can choose A so that $A(S)$ is the graph of one of the three functions

$$(b1) \quad (s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{C}{6}(s^3 + 3st^2)$$

$$(b2) \quad (s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{C}{6}(t^3 + 3ts^2)$$

$$(c) \quad (s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{1}{6}(s + t)^3;$$

we can also choose A so that $A(S)$ is the graph of one of the two functions

$$(b') \quad (s, t) \mapsto st + \frac{D}{6}(s^3 + t^3)$$

$$(c') \quad (s, t) \mapsto st + \frac{1}{6}s^3.$$

Remark: Changing from (b1) to (b2) involves interchanging the first two axes of \mathbb{R}^3 while leaving the third axis fixed, so we need both forms unless we allow maps $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ of the form $A = T \circ B$ with T a translation and $\det B = \pm 1$. Alternatively, we can make do with only one of these forms if we allow ourselves to change the orientation of M , and hence the direction of ν_p .

PROOF. We might as well assume that $p = 0 \in \mathbb{R}^3$, the tangent plane M_p is the (x, y) -plane, the affine normal ν_p is $(0, 0, 1)$, and the vectors $(1, 0)$ and $(0, 1)$ are an orthonormal basis of M_p ; for it is clear that the image of M under a suitable special linear affine map will have this property. Then M is the image of an immersion $f(s, t) = (s, t, h(s, t))$ with $h(0, 0) = h_1(0, 0) = h_2(0, 0) = 0$; since the z -axis is the affine normal direction, the apolarity conditions become

$$(*) \quad \begin{cases} h_{22}h_{111} - 2h_{12}h_{112} + h_{11}h_{122} = 0 \\ h_{22}h_{112} - 2h_{12}h_{122} + h_{11}h_{222} = 0 \end{cases} \quad \text{at } (0, 0).$$

Suppose that p is an elliptic point. The fact that $\nu_p = (0, 0, 1)$ and the basis $(1, 0, 0), (0, 1, 0)$ is orthonormal, means that $h_{ij}(0, 0) = \delta_{ij}$. So $(*)$ becomes

$$h_{111} + h_{122} = 0 \quad \text{and} \quad h_{112} + h_{222} = 0 \quad \text{at } (0, 0),$$

and S is the graph of a function of the form

$$(s, t) \mapsto \frac{1}{2}(s^2 + t^2) + \frac{1}{6}(as^3 + 3bs^2t - 3ast^2 - bt^3).$$

Now suppose we apply one more special linear (in fact, special orthogonal) transformation

$$(1) \quad (\sigma, \tau, z) \mapsto (\sigma \cos \theta - \tau \sin \theta, \sigma \sin \theta + \tau \cos \theta, z).$$

The image of S under this map is the graph of

$$(s, t) \mapsto \frac{1}{2}([s \cos \theta - t \sin \theta]^2 + [s \sin \theta + t \cos \theta]^2) \\ + \frac{1}{6}(a[s \cos \theta - t \sin \theta]^3 + \cdots),$$

which works out to be

$$(2) \quad (s, t) \mapsto \frac{1}{2}(s^2 + t^2) + \frac{1}{6}(a^*s^3 + 3b^*s^2t - 3a^*st^2 - b^*t^3),$$

where

$$(3) \quad \begin{cases} a^* = a \cos^3 \theta + 3b \cos^2 \theta \sin \theta - 3a \cos \theta \sin^2 \theta - b \sin^3 \theta \\ b^* = b \cos^3 \theta - 3a \cos^2 \theta \sin \theta - 3b \cos \theta \sin^2 \theta + a \sin^3 \theta. \end{cases}$$

To obtain the desired form (a), we just have to choose θ so that $b^* = 0$, which can be done by choosing θ so that $\cot \theta$ is a solution of the equation

$$b(\cot \theta)^3 - 3a(\cot \theta)^2 - 3b(\cot \theta) + a = 0.$$

Now suppose that p is a hyperbolic point. Then $h_{11}(0, 0) = -h_{22}(0, 0) = 1$ and $h_{12}(0, 0) = 0$. So (*) becomes

$$h_{122} = h_{111} \quad \text{and} \quad h_{222} = h_{112} \quad \text{at } (0, 0),$$

and S is the graph of a function of the form

$$(4) \quad (s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{1}{6}(as^3 + 3bs^2t + 3ast^2 + bt^3).$$

We might as well assume that $a, b \neq 0$, for otherwise we already have the form (b1) or (b2). We apply one more special linear transformation

$$(5) \quad (\sigma, \tau, z) \mapsto (\sigma \cosh u + \tau \sinh u, \sigma \sinh u + \tau \cosh u, z).$$

The image of S under this map works out to be the graph of

$$(s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{1}{6}(a^*s^3 + 3b^*s^2t + 3a^*st^2 + b^*t^3),$$

where

$$(6) \quad \begin{cases} a^* = a \cosh^3 u + 3b \cosh^2 u \sinh u + 3a \cosh u \sinh^2 u + b \sinh^3 u \\ b^* = b \cosh^3 u + 3a \cosh^2 u \sinh u + 3b \cosh u \sinh^2 u + a \sinh^3 u. \end{cases}$$

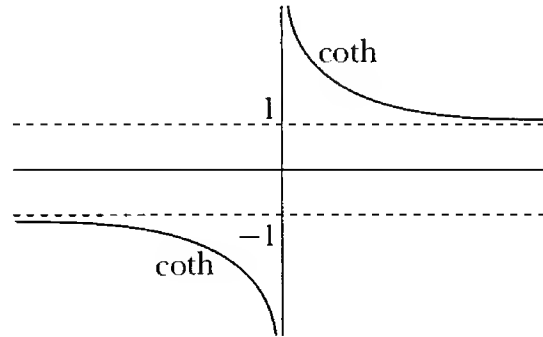
To obtain the form (b1), we have to find u so that $v = \coth u = (\cosh u)/(\sinh u)$ satisfies the equation

$$(7) \quad bv^3 + 3av^2 + 3bv + a = 0.$$

Now \coth does not take on all values, for the odd function

$$\coth u = \frac{\cosh u}{\sinh u} = \frac{e^u + e^{-u}}{e^u - e^{-u}} = \frac{e^{2u} + 1}{e^{2u} - 1}$$

is clearly > 1 for $u > 0$, and hence < -1 for $u < 0$. It is easy to see that \coth actually takes on all values except those in $[-1, 1]$. So we can obtain the



form (b1) if equation (7) has a real root which is not in $[-1, 1]$. This certainly happens if the values of $bv^3 + 3av^2 + 3bv + a$ are either both positive or both negative at $v = 1$ and $v = -1$. So we can obtain (b1) if

$$(8) \quad 4a + 4b \text{ and } 4a - 4b \text{ are both } > 0 \text{ or both } < 0.$$

But similarly, we can obtain (b2) if the values of $av^3 + 3bv^2 + 3av + b$ are both positive or both negative at $v = 1$ and $v = -1$, i.e., if

$$(9) \quad 4a + 4b \text{ and } 4b - 4a \text{ are both } > 0 \text{ or both } < 0.$$

Clearly, either (8) or (9) must hold if $a \neq \pm b$. Thus (b1) or (b2) can be obtained when $a \neq \pm b$. We note right away that (b') is obtained from (b1) by considering the special linear transformation

$$(10) \quad (\sigma, \tau, z) \mapsto \left(\frac{\sigma + \tau}{\sqrt{2}}, \frac{-(\sigma - \tau)}{\sqrt{2}}, z \right),$$

the constant D being $-C\sqrt{2}$. Similarly, (b') is obtained from (b2) by considering the special linear transformation

$$(11) \quad (\sigma, \tau, z) \mapsto \left(\frac{\sigma - \tau}{\sqrt{2}}, \frac{\sigma + \tau}{\sqrt{2}}, z \right),$$

the constant D being $+C\sqrt{2}$.

When we have $a = \pm b$, then our original function (4) is

$$(s, t) \mapsto \frac{1}{2}(s^2 - t^2) + \frac{a}{6}(s \pm t)^3.$$

The transformation (11) or (10) will change this to

$$(s, t) \mapsto st + \frac{a'}{6}s^3.$$

If $a' = 0$, this is a special case of (b'). Otherwise, we can make $a' = 1$, and thus achieve (c'), by a transformation which changes s to $s/\sqrt[3]{a'}$, and t to $\sqrt[3]{a'}t$. Since we can achieve (c'), we can now achieve (c) by using the inverse of the transformation (11). ♦

Notice that in case (a), the cubic part of S is just (a multiple of) the monkey saddle. Implicit in the previous Proposition is a geometric interpretation of apolarity when $\Phi: V \rightarrow \mathbb{R}$ corresponds to a positive definite inner product on a 2-dimensional vector space V . The cubic form $\Psi: V \rightarrow \mathbb{R}$ is apolar to Φ when the set $\Psi^{-1}(0)$ consists of three lines that make angles of $\pi/3$ with each other (in the inner product on V corresponding to Φ). It is clear that if we apply a rotation around the z -axis through an angle of $\pi/3$ or $2\pi/3$, then $A(S)$ will go into itself, so in case (a) the affine linear map A is not unique. We can also change C to $-C$ in (a) by rotation through an angle of π , which changes s to $-s$ and t to $-t$; the same change occurs if we rotate through an angle of $\pi + \pi/3$ or $\pi + 2\pi/3$. If $C = 0$, then we can apply any rotation around the z -axis. However, from equations (3) it is easy to see that this is the only extent to which A and C are not unique.

In the hyperbolic case, if we can obtain (b1) or (b2) with one constant $C \neq 0$, then we can also achieve this form with $-C$ by applying the rotation $(\sigma, \tau, z) \mapsto (-\sigma, -\tau, z)$. On the other hand, it is easy to see that this is the only other constant we can obtain, and that there is precisely one A which will make $A(S)$ have each of these forms. When $C = 0$, then we can always compose A with any transformation of the form

$$(\sigma, \tau, z) \mapsto (\pm[\sigma \cosh u + \tau \sinh u], \pm[\sigma \sinh u + \tau \cosh u], z).$$

Bearing these remarks in mind, we see that Proposition 17 allows us to define a single new invariant: If p is a point of M for which the osculating cubic S has the form (a), (b1), or (b2) for some number C (unique up to sign), we define the **Pick invariant** J at p by

$$J = \frac{C^2}{2};$$

if the osculating cubic has the form (c), we define J to be 0.

It would be nice to have a straightforward invariant description of J in terms of the forms \mathbf{I} , \mathbf{II} , and it can be obtained as follows. Given an inner product $\langle \cdot, \cdot \rangle$ on a vector space V , we can use it to define an inner product $\langle \cdot, \cdot \rangle$ on the vector space of all trilinear maps $\alpha: V \times V \times V \rightarrow \mathbb{R}$. Instead of describing this completely invariantly, let us consider a basis v_1, \dots, v_n of V and let $\gamma_{ij} = \langle v_i, v_j \rangle$. The matrix (γ_{ij}) is non-singular, and has an inverse matrix (γ^{ij}) . Given two trilinear maps $\alpha, \beta: V \times V \times V \rightarrow \mathbb{R}$, let $\alpha_{ijk} = \alpha(v_i, v_j, v_k)$ and $\beta_{ijk} = \beta(v_i, v_j, v_k)$. Then

$$\langle \alpha, \beta \rangle = \sum_{i,j,k=1}^n \sum_{\rho,\sigma,\tau=1}^n \alpha_{ijk} \beta_{\rho\sigma\tau} \gamma^{i\rho} \gamma^{j\sigma} \gamma^{k\tau}.$$

If v_1, \dots, v_n is orthonormal with respect to $\langle \cdot, \cdot \rangle$, then we have simply

$$\langle \alpha, \beta \rangle = \sum_{i,j,k=1}^n \alpha_{ijk} \beta_{ijk}.$$

The reader may easily check that this definition does not depend on the choice of basis, or else fashion an invariant definition.

18. PROPOSITION. For a point p of a surface $M \subset \mathbb{R}^3$, the Pick invariant $J(p)$ is

$$J(p) = \frac{1}{2} \langle \mathbf{II}(p), \mathbf{II}(p) \rangle_p,$$

where $\langle \cdot, \cdot \rangle_p$ is the inner product on all trilinear maps on M_p determined by the inner product $\langle \cdot, \cdot \rangle_p$ on M_p .

If $f: U \rightarrow \mathbb{R}^3$ is an immersion, then

$$J = \frac{1}{2} \sum_{i,j,k=1}^2 \sum_{\rho,\sigma,\tau=1}^2 f_{ijk} f_{\rho\sigma\tau} g^{i\rho} g^{j\sigma} g^{k\tau}.$$

PROOF. It suffices to consider the case where $M = \{(s, t, h(s, t))\}$, the point p is $0 \in \mathbb{R}^3$, and the osculating cubic S has one of the forms (a)–(c) in Proposition 17. In case (a) we thus have

$$\begin{aligned} h(0, 0) &= h_i(0, 0) = 0, \\ h_{ij}(0, 0) &= \delta_{ij}, \\ h_{111}(0, 0) &= C, \\ h_{211}(0, 0) &= h_{121}(0, 0) = h_{112}(0, 0) = -C, \quad \text{all other } h_{ijk}(0, 0) = 0, \end{aligned}$$

with similar equations in cases (b1), (b2), (c). We can compute the $g_{ij}(0, 0)$ from equation (I) on page 90 and the $\ell_{ijk}(0, 0)$ from equation (3) on page 110. It is then a simple calculation to check that we do indeed have

$$\begin{aligned} J(p) &= \frac{1}{2} \sum_{i,j,k=1}^2 \sum_{\rho,\sigma,\tau=1}^2 \ell_{ijk} \ell_{\rho\sigma\tau} g^{i\rho} g^{j\sigma} g^{k\tau} \quad \text{at } (0, 0) \\ &= \frac{1}{2} \{\mathfrak{I}(p), \mathfrak{I}(p)\}_p. \end{aligned}$$

Since both sides of our formula have an invariant meaning, the formula must be true for any coordinate system. ♦

Notice also that for a moving frame X_1, X_2 as on page 106 we have

$$J = \frac{1}{2} \sum_{i,j,k} (c_{ijk})^2$$

in the elliptic case, with a similar formula in the hyperbolic case.

Now that we have obtained the invariant J , there is an obvious question staring us in the face: what are the surfaces with $J = 0$ everywhere? In ordinary surface theory we found that surfaces with $k_1 = k_2 = 0$ everywhere are planes. It seems natural to conjecture that the surfaces with $J = 0$ everywhere are just the surfaces which can be described by quadratic equations. Now this isn't true, and the reason is essentially because we have $J = 0$ in case (c) of Proposition 17. In cases (a), (b1), (b2), the vanishing of $J(p)$ implies the vanishing of $\mathfrak{I}(p)$ [as is also clear in case (a) from Proposition 18]. Leaving the complexities which arise because of case (c) to later Problems (4-14, 15), we restrict our attention to surfaces with $\mathfrak{I} = 0$ everywhere.

19. PROPOSITION. If $M \subset \mathbb{R}^3$ is a connected surface with all points elliptic or all points hyperbolic, and $\mathfrak{I} = 0$ on M , then M is a quadratic surface, that is, $M = W^{-1}(0)$ for some $W: \mathbb{R}^3 \rightarrow \mathbb{R}$ of the form

$$W(x_1, x_2, x_3) = \sum_{i,j} a_{ij} x_i x_j + \sum_i b_i x_i + c.$$

(These surfaces are described in greater detail in the next chapter; at the moment it is only necessary to note that the affine linear image of a quadratic surface is also a quadratic surface.)

Conversely, every non-flat quadratic surface has $\mathfrak{I} = 0$ everywhere.

PROOF. As usual, we consider only the elliptic case. Take a moving frame X_1, X_2, X_3 on M as on page 106. If $\mathfrak{I} = 0$, then all $c_{ijk} = 0$, so equation (I) on page 106 shows that

$$\psi_j^i = \omega_j^i,$$

and in particular

$$\psi_j^i = -\psi_i^j.$$

This implies that

$$d\psi_j^i + \sum_{k=1}^2 \psi_k^i \wedge \psi_j^k = -\left(d\psi_i^j + \sum_{k=1}^2 \psi_k^j \wedge \psi_i^k\right).$$

But

$$\begin{aligned} \text{(I)} \quad d\psi_j^i &= -\sum_{k=1}^2 \psi_k^i \wedge \psi_j^k - \psi_3^i \wedge \psi_j^3 \\ &= -\sum_{k=1}^2 \psi_k^i \wedge \psi_j^k - \psi_3^i \wedge \theta^j, \end{aligned}$$

so we obtain

$$\psi_3^i \wedge \theta^j = -(\psi_3^j \wedge \theta^i).$$

Taking $i = j$, we see that ψ_3^i is a multiple of θ^i , and taking $i \neq j$ we see that the multiples are the same for each i , so we have

$$\text{(2)} \quad \psi_3^i = \alpha \cdot \theta^i$$

for some function α . Taking d of equation (2) gives

$$\begin{aligned} -\sum_{k=1}^2 \psi_k^i \wedge \psi_3^k &= d\psi_3^i = d\alpha \wedge \theta^i + \alpha d\theta^i \\ &\Downarrow \\ -\sum_{k=1}^2 \psi_k^i \wedge \alpha \cdot \theta^k &= d\alpha \wedge \theta^i - \alpha \left(\sum_{k=1}^2 \psi_k^i \wedge \theta^k \right), \end{aligned}$$

and hence

$$d\alpha \wedge \theta^i = 0.$$

Thus $d\alpha = 0$, and α is a constant. We consider two cases.

Case 1. $\alpha = 0$. Equation (2) shows that $\psi_3^i = 0$, so for tangent vectors X on M we have

$$\nabla'_X v = \nabla'_X X_3 = \sum_{i=1}^2 \psi_3^i(X) \cdot X_i = 0,$$

i.e., v is constant; we will use it as one of our axes. We also have, by (1),

$$d\psi_j^i = -\sum_{k=1}^2 \psi_k^i \wedge \psi_j^k.$$

Since $\psi_j^i = \omega_j^i$, which are the connection forms for the metric $\langle \cdot, \cdot \rangle$, this equation shows that $\langle \cdot, \cdot \rangle$ is *flat*. So we could have chosen our orthonormal moving frame X_1, X_2 to be of the form $X_i = f_i$ for some isometry $f: U \rightarrow \mathbb{R}^3$, where $U \subset \mathbb{R}^2$ has the standard Riemannian metric; and with this choice we have $\psi_j^i = \omega_j^i = 0$. Now we have

$$\begin{aligned} f_{ij} = \nabla'_{f_j} f_i &= \sum_{k=1}^2 \psi_i^k(f_j) \cdot f_k + \psi_i^3(f_j) \cdot v \\ &= 0 + \theta^i(f_j)v = \delta_{ij}v. \end{aligned}$$

So f is of the form

$$f(s, t) = c + b_1 s + b_2 t + \frac{1}{2}(s^2 + t^2)v$$

for constants b_1, b_2, c .

Case 2. $\alpha \neq 0$. Now equation (2) shows that

$$\nabla'_X v = \sum_{i=1}^2 \alpha \theta^i(X) \cdot X_i = \alpha X.$$

So for any $f: U \rightarrow M \subset \mathbb{R}^3$ we have

$$\mathcal{N}_i = \alpha f_i \implies \mathcal{N} = \alpha f + \beta \implies f = \frac{1}{\alpha}(\mathcal{N} - \beta)$$

for some constant vector β . Hence it suffices to show that v satisfies a quadratic equation.

Let \mathbf{X} be the 3×3 matrix

$$\mathbf{X} = (X_1, X_2, X_3) = (X_1, X_2, v)$$

where the X_i are considered as column vectors. Then $d\mathbf{X}$ is given in terms of the 3×3 matrix $\psi = (\psi_\beta^\alpha)$ as

$$(1) \quad d\mathbf{X} = \mathbf{X} \cdot \psi.$$

Since ψ is actually

$$\psi = \begin{pmatrix} \psi_1^1 & \psi_2^1 & \alpha\theta^1 \\ \psi_1^2 & \psi_2^2 & \alpha\theta^2 \\ \theta^1 & \theta^2 & 0 \end{pmatrix}, \quad \psi_j^i = -\psi_i^j,$$

we easily see that if we set

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\alpha} \end{pmatrix},$$

then

$$(2) \quad \psi \cdot A + A \cdot \psi^t = 0.$$

Thus we have

$$\begin{aligned} d(\mathbf{X} \cdot A \cdot \mathbf{X}^t) &= d\mathbf{X} \cdot A \cdot \mathbf{X}^t + \mathbf{X} \cdot A \cdot d\mathbf{X}^t \\ &= \mathbf{X} \cdot \psi \cdot A \cdot \mathbf{X}^t + \mathbf{X} \cdot A \cdot \psi^t \cdot \mathbf{X}^t && \text{by (1)} \\ &= 0 && \text{by (2)}. \end{aligned}$$

We can assume that \mathbf{X} is the identity matrix at some point, so we obtain

$$\begin{aligned} \mathbf{X} \cdot A \cdot \mathbf{X}^t &= A \implies (A^{-1} \cdot \mathbf{X} \cdot A) \cdot \mathbf{X}^t = \text{identity} \\ &\implies \mathbf{X}^t \cdot (A^{-1} \cdot \mathbf{X} \cdot A) = \text{identity} \\ &\implies \mathbf{X}^t \cdot A^{-1} \cdot \mathbf{X} = A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -\alpha \end{pmatrix}. \end{aligned}$$

The (3, 3) entry of this equation gives

$$\mathbf{v} \cdot A^{-1} \cdot \mathbf{v}^t = -\alpha,$$

which is a quadratic equation for \mathbf{v} .

It is not hard to show, conversely, that all surfaces of the sort we have obtained have $\mathfrak{I} = 0$. These are the only non-flat quadrics (see Problem 3-6). ♦

We proceed in our study of special affine surface theory along the very same route followed in ordinary surface theory, by looking for the equations expressing the derivatives of a moving frame X_1, X_2, \mathbf{v} in terms of this frame, and by looking at the integrability conditions for these equations. We will always work only with the elliptic case, leaving the hyperbolic case to the reader. We already have the analogue of the Gauss formulas,

$$\nabla'_X Y = \nabla_X Y + \mathfrak{A}(X, Y) + \langle X, Y \rangle \mathbf{v},$$

or in terms of a map $f: U \rightarrow \mathbb{R}^3$,

$$\begin{aligned} f_{ij} &= \nabla_{f_i} f_j + \mathfrak{A}(f_i, f_j) + g_{ij} \mathcal{N} \\ &= \sum_{k=1}^2 \Gamma_{ij}^k f_k + \sum_{i=1}^2 \ell_{ij}^k f_k + g_{ij} \mathcal{N}; \end{aligned}$$

in terms of a moving frame (X_1, X_2, \mathbf{v}) with X_1, X_2 orthonormal for $\langle \cdot, \cdot \rangle$ we have

$$\psi_i^j = \omega_i^j + \sum_{k=1}^2 c_{ijk} \theta^k, \quad \psi_i^3 = \theta^i.$$

The strange way in which the special affine Gauss formulas interchange the roles of $\mathfrak{T} \nabla'_X Y$ and $\mathfrak{L} \nabla'_X Y$, as compared to the ordinary Gauss formulas, will be reflected by funny twists throughout the development.

The next thing we seek is an analogue of the Weingarten equations. Now we know that

$$\nabla'_{X_p} \nu \in M_p,$$

so for a map $f: U \rightarrow \mathbb{R}^3$ we always have

$$\mathcal{N}_i = \sum_{j=1}^2 \mathfrak{b}_i^j f_j$$

for certain functions \mathfrak{b}_i^j with

$$\langle n_i, f_j \rangle = \sum_{\rho=1}^2 \mathfrak{b}_i^\rho g_{\rho j} = \mathfrak{b}_{ij}, \quad \text{say;}$$

for our moving frame we have

$$\nabla'_X \nu = \nabla'_X X_3 = \sum_{i=1}^2 \psi_3^i(X) \cdot X_i.$$

Unlike the situation in ordinary surface theory however, it is not at all clear how, or even whether, the \mathfrak{b}_i^j and \mathfrak{b}_{ij} are related to the ℓ_{ijk} . This is answered in a very unexpected way when we look at the integrability conditions for the special affine Gauss equations. These can be derived from the formulas for f_{ij} , exactly as in the first part of this chapter, or from the formulas for $\nabla'_X Y$, as in Chapter 1. But since the computations are quite involved, no matter how one does them, it will be easiest to use the moving frame version. We begin with the tangential part of these equations,

$$(1) \quad \psi_i^j = \omega_i^j + \sum_{k=1}^2 c_{ijk} \theta^k.$$

We take d of this equation, remembering that $\psi_i^3 = \theta^i$, and introducing the curvature forms Ω_i^j for the ω_i^j , to obtain

$$(2) \quad -\sum_{\rho=1}^2 \psi_\rho^j \wedge \psi_i^\rho - \psi_3^j \wedge \theta^i = -\sum_{\rho=1}^2 \omega_\rho^j \wedge \omega_i^\rho + \Omega_i^j \\ + \sum_{k=1}^2 dc_{ijk} \wedge \theta^k + \sum_{k=1}^2 c_{ijk} d\theta^k.$$

Substituting back from (1) we have

$$\begin{aligned} & - \sum_{\rho} \left\{ \omega_{\rho}^j + \sum_k c_{\rho j k} \theta^k \right\} \wedge \left\{ \omega_i^{\rho} + \sum_k c_{i \rho k} \theta^k \right\} - \psi_3^j \wedge \theta^i \\ & = - \sum_{\rho} \omega_{\rho}^j \wedge \omega_i^{\rho} + \Omega_i^j + \sum_k d c_{i j k} \wedge \theta^k - \sum_{k, \rho} c_{i j k} \omega_{\rho}^k \wedge \theta^{\rho}. \end{aligned}$$

Using $\omega_i^{\rho} = -\omega_{\rho}^i$ and switching dummy indices in the last term, we get

$$\begin{aligned} (3) \quad & - \sum_{\rho, k, l} c_{\rho j k} c_{i \rho l} \theta^k \wedge \theta^l - \psi_3^j \wedge \theta^i \\ & = \Omega_i^j + \sum_k \left[d c_{i j k} - \sum_{\rho} c_{\rho j k} \omega_i^{\rho} - \sum_{\rho} c_{i \rho k} \omega_j^{\rho} - \sum_{\rho} c_{i j \rho} \omega_k^{\rho} \right] \wedge \theta^k. \end{aligned}$$

To interpret equation (3), we apply it to two vectors, X, Y . The first term becomes

$$\begin{aligned} (3a) \quad & \sum_{\rho, k, l} c_{j k \rho} c_{i l \rho} \theta^k(Y) \theta^l(X) - \sum_{\rho, k, l} c_{j k l} c_{i l \rho} \theta^k(X) \theta^l(Y) \\ & = \langle \mathfrak{A}(X_i, X), \mathfrak{A}(X_j, Y) \rangle - \langle \mathfrak{A}(X_i, Y), \mathfrak{A}(X_j, X) \rangle. \end{aligned}$$

The second term becomes

$$\begin{aligned} (3b) \quad & \psi_3^j(Y) \theta^i(X) - \psi_3^j(X) \theta^i(Y) = \langle d v(Y), X_j \rangle \cdot \langle X, X_i \rangle \\ & \quad - \langle d v(X), X_j \rangle \cdot \langle Y, X_i \rangle, \end{aligned}$$

while the first term on the right becomes simply

$$(3c) \quad \langle \mathcal{R}(X, Y) X_i, X_j \rangle,$$

where \mathcal{R} is the curvature tensor for $\langle \cdot, \cdot \rangle$. To interpret the last term, we recall that the tensor

$$\mathfrak{I} = \sum c_{i j k} \theta^i \otimes \theta^j \otimes \theta^k$$

has a covariant derivative $\nabla \mathfrak{I}(X, Y, Z, W) = (\nabla_W \mathfrak{I})(X, Y, Z)$, which can be written

$$\nabla \mathfrak{I} = \sum_{i, j, k, l} c_{i j k, l} \theta^i \otimes \theta^j \otimes \theta^k \otimes \theta^l, \quad \text{say.}$$

Using Corollary I.6-5 one easily checks (Problem 5) that

$$\sum_l c_{ijk;l} \theta^l = dc_{ijk} - \sum_\rho c_{\rho jk} \omega_i^\rho - \sum_\rho c_{i\rho k} \omega_j^\rho - \sum_\rho c_{ij\rho} \omega_k^\rho.$$

So the last term is

$$\sum_{k,l} c_{ijk;l} \theta^l \wedge \theta^k,$$

which when applied to X, Y gives

$$\begin{aligned} (3d) \quad \sum_{k,l} c_{ijk;l} \theta^l(X) \theta^k(Y) - \sum_{k,l} c_{ijk;l} \theta^l(Y) \theta^k(X) \\ = (\nabla_X \mathbb{I})(X_i, X_j, Y) - (\nabla_Y \mathbb{I})(X_i, X_j, X). \end{aligned}$$

Writing (3a) + (3b) = (3c) + (3d), but replacing X_i, X_j by arbitrary tangent vectors Z, W , we thus obtain

$$\begin{aligned} (A) \quad \langle \mathcal{R}(X, Y)Z, W \rangle &= (\nabla_Y \mathbb{I})(W, Z, X) - (\nabla_X \mathbb{I})(W, Z, Y) \\ &\quad + \langle \mathfrak{s}(W, Y), \mathfrak{s}(X, Z) \rangle - \langle \mathfrak{s}(W, X), \mathfrak{s}(Y, Z) \rangle \\ &\quad + \langle d\mathfrak{v}(Y), W \rangle \cdot \langle X, Z \rangle - \langle d\mathfrak{v}(X), W \rangle \cdot \langle Y, Z \rangle. \end{aligned}$$

In terms of a map $f: U \rightarrow \mathbb{R}^3$ this becomes

$$\begin{aligned} \mathcal{R}_{jik\mu} &= \ell_{jik;\mu} - \ell_{ji\mu;k} + \sum_\rho (\ell_{ik}^\rho \ell_{j\rho\mu} - \ell_{i\mu}^\rho \ell_{j\rho k}) \\ &\quad + g_{ik} b_{\mu j} - g_{i\mu} b_{kj}. \end{aligned}$$

Now what do these equations tell us? In ordinary surface theory we obtained the Theorema Egregium, telling us that the intrinsic Gaussian curvature K is equal to some expression involving the coefficients of \mathbb{II} . But now we don't get anything of the sort, because we have the unknown expressions* $d\mathfrak{v}$ (or b_{ij}). Instead, equation (A) allows us to solve for $\langle d\mathfrak{v}(Y), W \rangle$ —we just have to choose a unit vector Z with $\langle Y, Z \rangle = 0$, and set $X = Z$.

*Note also that the terms involving \mathfrak{s} in equation (A) are the negatives of the corresponding terms involving \mathfrak{s} in the ordinary Gauss equations.

An especially nice formula for $d\nu$ can be obtained with a little more work. Recall first of all that for a map $f: U \rightarrow \mathbb{R}^3$ we have the apolarity conditions

$$\begin{aligned} 0 = \sum_{i,j} g^{ij} \ell_{\mu ij} &\implies 0 = \sum_{i,j} (g^{ij} \ell_{\mu ij})_{;k} \\ &= \sum_{i,j} g^{ij} \ell_{\mu ij;k}, \quad \text{by Ricci's Lemma.} \end{aligned}$$

For orthonormal X_1, X_2 this means that

$$\sum_i (\nabla_X \mathfrak{I})(X_i, X_i, Y) = 0.$$

[Naturally, this formula can also be derived directly from the apolarity condition $\sum_i \mathfrak{I}(X_i, X_i, Y) = 0$, but the coordinate treatment is much easier and quicker.] So if \mathcal{K} is the intrinsic Gaussian curvature for the metric $\langle \cdot, \cdot \rangle$, then equation (A) gives

$$\begin{aligned} -2\mathcal{K} &= \sum_{i,j} \langle \mathcal{R}(X_i, X_j) X_i, X_j \rangle \\ &= \sum_j \left\{ \sum_i (\nabla_{X_j} \mathfrak{I})(X_j, X_i, X_i) \right\} - \sum_i \left\{ \sum_j (\nabla_{X_i} \mathfrak{I})(X_j, X_i, X_i) \right\} \\ &\quad + \sum_{i,j} \langle \mathfrak{A}(X_j, X_j), \mathfrak{A}(X_i, X_i) \rangle - \sum_{i,j} \langle \mathfrak{A}(X_i, X_j), \mathfrak{A}(X_i, X_j) \rangle \\ &\quad + \sum_{i,j} \langle d\nu(X_j), X_j \rangle \cdot \delta_{ii} - \sum_{i,j} \langle d\nu(X_i), X_j \rangle \cdot \delta_{ij} \\ &= 0 - 0 + \sum_{i,j,k} c_{jjk} c_{iik} - \sum_{i,j,k} (c_{ijk})^2 \\ &\quad + 2 \sum_j \langle d\nu(X_j), X_j \rangle - \sum_i \langle d\nu(X_i), X_i \rangle \\ &= \sum_{j,k} c_{jjk} \left(\sum_i c_{iik} \right) - \sum_{i,j,k} (c_{ijk})^2 + \sum_j \langle d\nu(X_j), X_j \rangle. \end{aligned}$$

Now $\sum_i c_{iik} = 0$ by the apolarity conditions, so we have (see page 117)

$$(1) \quad -2\mathcal{K} = -2J + \sum_j \langle d\nu(X_j), X_j \rangle.$$

where J is the Pick invariant. But as another consequence of equation (A), and the symmetry properties of \mathcal{R} , we also have

$$\begin{aligned}
 (2) \quad 0 &= \sum_i \langle \mathcal{R}(X_i, X) X_i, Y \rangle + \langle \mathcal{R}(X_i, X) Y, X_i \rangle \\
 &= 0 - 2 \sum_i (\nabla_{X_i} \mathfrak{I})(X, Y, X_i) \\
 &\quad + \sum_i \langle d\mathfrak{v}(X), Y \rangle \cdot \delta_{ii} - \sum_i \langle d\mathfrak{v}(X_i), Y \rangle \cdot \langle X, X_i \rangle \\
 &\quad + \sum_i \langle d\mathfrak{v}(X), X_i \rangle \cdot \langle X_i, Y \rangle - \sum_i \langle d\mathfrak{v}(X_i), X_i \rangle \cdot \langle X, Y \rangle \\
 &= -2 \sum_i (\nabla_{X_i} \mathfrak{I})(X, Y, X_i) + 2 \langle d\mathfrak{v}(X), Y \rangle \\
 &\quad - \sum_i \langle d\mathfrak{v}(X_i), X_i \rangle \cdot \langle X, Y \rangle.
 \end{aligned}$$

To interpret the sum $\sum_i (\nabla_{X_i} \mathfrak{I})(X, Y, X_i)$, we again switch to coordinates for simplicity, noting to begin with that the sum is easily shown to be independent of the particular orthonormal X_1, X_2 chosen. Now this sum is just the value of

$$\sum_{i,j} g^{ij} \ell_{k\mu i;j} = \sum_j \ell_{k\mu;j}^j$$

when we happen to have $X_i = f_i$ ($i = 1, 2$) and $f_k = X$, $f_\mu = Y$. Since $\sum_j \ell_{k\mu;j}^j$ are the components of a tensor, namely the tensor

$$\mathfrak{J}(X, Y) = \text{trace}(Z \mapsto (\nabla_Z \mathfrak{J})(X, Y)),$$

we must, in fact, always have

$$\sum_i (\nabla_{X_i} \mathfrak{I})(X, Y, X_i) = \mathfrak{J}(X, Y).$$

Thus equation (2) can be written

$$(3) \quad 2 \langle d\mathfrak{v}(X), Y \rangle = \sum_i \langle d\mathfrak{v}(X_i), X_i \rangle \cdot \langle X, Y \rangle + 2 \mathfrak{J}(X, Y).$$

Substituting in from (1) we obtain

$$(4) \quad \langle d\mathfrak{v}(X), Y \rangle = (J - \mathcal{K}) \cdot \langle X, Y \rangle + \mathfrak{J}(X, Y),$$

the analogue in ordinary surface theory being

$$\langle d\nu(X), Y \rangle = -K \cdot \langle X, Y \rangle.$$

We can solve equation (4) for $d\nu(X)$ explicitly by introducing the tensor $\tilde{\mathcal{G}}$ of type $\binom{1}{1}$ defined by

$$\langle \tilde{\mathcal{G}}(X), Y \rangle = \mathcal{G}(X, Y).$$

Then we have

$$(B) \quad d\nu(X) = (J - \mathcal{K}) \cdot X + \tilde{\mathcal{G}}(X).$$

In terms of a map $f: U \rightarrow \mathbb{R}^3$ we have

$$\begin{aligned} \mathfrak{L}_{\mu k} &= g_{\mu k} \cdot (J - \mathcal{K}) + \sum_j \ell_{\mu k; j}^j \\ &= g_{\mu k} \cdot (J - \mathcal{K}) + \mathcal{L}_{\mu k}, \quad \text{say;} \end{aligned}$$

in terms of our moving frame we have

$$\begin{aligned} \psi_3^j &= (J - \mathcal{K})\theta^j + \sum_{i, k} c_{ijk; k} \theta^i \\ &= (J - \mathcal{K})\theta^j + \sum_i C_{ij} \theta^i, \quad \text{say.} \end{aligned}$$

Conversely, it is not hard to see that if we define $d\nu(X)$ [or the $\mathfrak{L}_{\mu k}$, or the ψ_3^j] by these formulas, then equation (A) [or the coordinate equation right below it, or equation (3) on page 123] is satisfied.

As one immediate consequence of equation (4) we find that the special affine normal ν has another property in common with the ordinary normal ν :

20. PROPOSITION. The map $d\nu: M_p \rightarrow M_p$ is self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle_p$ on M_p ,

$$\langle d\nu(X_p), Y_p \rangle = \langle X_p, d\nu(Y_p) \rangle \quad \text{for } X_p, Y_p \in M_p.$$

PROOF. We just need to show that $\mathcal{G}(X, Y) = \mathcal{G}(Y, X)$. This follows from the symmetry of \mathcal{G} , and the definition of \mathcal{G} (on page 126). ♦

The eigenvectors of $-dv: M_p \rightarrow M_p$ are naturally called the **special affine principal directions** at p , and the corresponding eigenvalues are the **special affine principal curvatures** k_1 and k_2 . The **special affine mean curvature** \mathcal{H} and **special affine (extrinsic) curvature** \mathcal{K}_{ext} are then defined by

$$\begin{aligned}\mathcal{H} &= \frac{1}{2}(k_1 + k_2) \\ \mathcal{K}_{\text{ext}} &= k_1 \cdot k_2.\end{aligned}$$

If $X_1, X_2 \in M_p$ are orthonormal, then

$$\mathcal{H}(p) = -\frac{1}{2} \sum_{j=1}^2 \langle dv(X_j), X_j \rangle.$$

So equation (1) on page 125 shows that we also have

$$\mathcal{H} = \mathcal{K} - J.$$

We have obtained these results by looking at the integrability conditions for the tangential part of the Gauss formulas. Now we will look at the integrability conditions for the \mathcal{L} component,

$$\psi_i^3 = \theta^i.$$

Exterior differentiation gives

$$-\sum_{k=1}^2 \psi_k^3 \wedge \psi_i^k = d\theta^i = -\sum_{k=1}^2 \psi_k^i \wedge \theta^k$$

or

$$-\sum_k \theta^k \wedge \psi_i^k = \sum_k \theta^k \wedge \psi_k^i \quad \text{or} \quad \sum_k (\psi_i^k + \psi_k^i) \wedge \theta^k = 0.$$

But these conditions are *automatic*, for we derived this equation in the proof of Proposition 16 (conversely, the equation follows immediately from equation (2) on page 106 and symmetry of the c_{ijk}).

On the other hand, we still have to look at the conditions which say that $\mathcal{N}_{ij} = \mathcal{N}_{ji}$. In ordinary surface theory they reduced to the Codazzi-Mainardi equations; now we will obtain new conditions. The moving frame version of the formula for \mathcal{N}_i is that on page 127.

$$(1) \quad \psi_3^j = (J - \mathcal{K})\theta^j + \sum_i C_{ij}\theta^i.$$

Exterior differentiation gives

$$\begin{aligned} - \sum_{\rho} \psi_{\rho}^j \wedge \psi_3^{\rho} &= d(J - \mathcal{K}) \wedge \theta^j + (J - \mathcal{K})d\theta^j + \sum_i dC_{ij} \wedge \theta^i \\ &\quad - \sum_i C_{ij} \wedge \left(\sum_{\rho} \omega_{\rho}^i \wedge \theta^{\rho} \right). \end{aligned}$$

Substituting in for ψ_3^k from (1), noting that $-\sum_{\rho} \psi_{\rho}^j \wedge \theta^{\rho} = d\theta^j$, and switching dummy indices in the last term on the right, we have

$$- \sum_{i,\rho} C_{i\rho} \psi_{\rho}^j \wedge \theta^i = d(J - \mathcal{K}) \wedge \theta^j + \sum_i dC_{ij} \wedge \theta^i - \sum_{i,\rho} C_{\rho j} \wedge \omega_i^{\rho} \wedge \theta^i.$$

Now writing the ψ_{ρ}^j in terms of the ω_{ρ}^j we get

$$\begin{aligned} - \sum_{i,\rho} C_{i\rho} \omega_{\rho}^j \wedge \theta^i &- \sum_{i,\rho,k} C_{i\rho} c_{j\rho k} \theta^k \wedge \theta^i \\ &= d(J - \mathcal{K}) \wedge \theta^j + \sum_i dC_{ij} \wedge \theta^i - \sum_{i,\rho} C_{\rho j} \wedge \omega_i^{\rho} \wedge \theta^i. \end{aligned}$$

Finally, since $\omega_{\rho}^j = -\omega_j^{\rho}$, we can write

$$\begin{aligned} (2) \quad &- \sum_{i,\rho,k} C_{i\rho} c_{j\rho k} \theta^k \wedge \theta^i \\ &= d(J - \mathcal{K}) \wedge \theta^j + \sum_i \left[dC_{ij} - \sum_{\rho} C_{\rho j} \wedge \omega_i^{\rho} - \sum_{\rho} C_{i\rho} \omega_j^{\rho} \right] \wedge \theta^i. \end{aligned}$$

To interpret this equation, we first apply it to (X, X_j) . The left side gives

$$(2a) \quad \sum_{i,\rho} C_{i\rho} c_{j\rho k} \theta^i(X) - \sum_{\rho,k} C_{j\rho} c_{j\rho k} \theta^k(X).$$

The first term on the right side gives

$$(2b) \quad X(J - \mathcal{K}) - X_j(J - \mathcal{K}) \cdot \theta^j(X).$$

For the other term on the right side we note, as on page 123, that the tensor $\mathcal{S} = \sum_{i,j} C_{ij} \theta^i \otimes \theta^j$ has a covariant derivative

$$\nabla \mathcal{S} = \sum_{i,j,k} C_{ij,k} \theta^i \otimes \theta^j \otimes \theta^k,$$

where (Problem 5)

$$\sum_k C_{ij;k} \theta^k = dC_{ij} - \sum_\rho C_{\rho j} \omega_i^\rho - \sum_\rho C_{i\rho} \omega_j^\rho.$$

So the second term on the right side of (2) is

$$\sum_{i,k} C_{ij;k} \theta^k \wedge \theta^i,$$

which when applied to (X, X_j) gives

$$(2c) \quad \sum_k C_{jj;k} \theta^k(X) - \sum_i C_{ij;j} \theta^i(X).$$

We now take the equations (2a) = (2b) + (2c) and add them for $j = 1, 2$ [the resultant equation is equivalent to the individual equations, for if a 2-form α satisfies $\sum_{j=1}^2 \alpha(X, X_j) = 0$, then also $\alpha(X_i, X_j) = 0$ for $i = 1, 2$], obtaining

$$(3) \quad \begin{aligned} & \sum_{i,\rho} C_{i\rho} \left(\sum_j c_{jj\rho} \right) \cdot \theta^i(X) - \sum_k \left(\sum_{j,\rho} C_{j\rho} c_{j\rho k} \right) \theta^k(X) \\ &= 2X(J - \mathcal{K}) - X(J - \mathcal{K}) \\ &+ \sum_k \left(\sum_j C_{jj;k} \right) \theta^k(X) - \sum_i \left(\sum_j C_{ij;j} \right) \theta^i(X). \end{aligned}$$

Now we have

$$(3a) \quad \sum_{i,\rho} C_{i\rho} \left(\sum_j c_{jj\rho} \right) \cdot \theta^i(X) = 0$$

by the apolarity conditions $\sum_j c_{jj\rho} = 0$. In order to interpret the term involving $\sum_{j,\rho} C_{j\rho} c_{j\rho k}$ we introduce the tensor $\mathfrak{J} * \mathfrak{I}$ of type $\binom{1}{0}$ defined by

$$\mathfrak{J} * \mathfrak{I}(X) = \sum_{i,j=1}^2 \mathfrak{J}(X_i, X_j) \cdot \mathfrak{I}(X, X_i, X_j),$$

where X_1, X_2 is any orthonormal basis; it is easily checked that this definition is independent of the choice of such basis. [This is the simplest description of $\mathfrak{J} * \mathfrak{I}$ —a completely invariant definition requires an orgy of linear algebra.

Classically, $\mathcal{J} * \mathfrak{I}$ is simply described in terms of its components, which are given by the 4-fold contraction

$$\sum_{i,\mu,\rho,\sigma} g^{\mu i} g^{\rho\sigma} \mathcal{L}_{\sigma i} \ell_{\mu\rho j},$$

where $\mathcal{L}_{\sigma i}$ are the components of \mathcal{J} .] Now we clearly have

$$(3b) \quad - \sum_k \left(\sum_{j,\rho} C_{j\rho} c_{j\rho k} \right) \theta^k(X) = -\mathcal{J} * \mathfrak{I}(X).$$

To deal with the term $\sum_j C_{jj;k}$ it is again simplest to switch to coordinates. We have the apolarity conditions

$$\begin{aligned} 0 &= \sum_{i,j} g^{ij} \ell_{ij\rho} \\ \implies 0 &= \sum_{i,j,\rho} g^{ij} g^{\mu\rho} \ell_{ij\rho} = \sum_{i,j} g^{ij} \ell_{ij}^{\mu} \\ \implies 0 &= \sum_{i,j,\mu} (g^{ij} \ell_{ij}^{\mu})_{;\mu} = \sum_{i,j,\mu} g^{ij} (\ell_{ij;\mu}^{\mu}) \quad \text{by Ricci's Lemma} \\ &= \sum_{i,j} g^{ij} \left(\sum_{\mu} \ell_{ij;\mu}^{\mu} \right) = \sum_{i,j} g^{ij} \mathcal{L}_{ij} \\ \implies 0 &= \sum_{i,j} (g^{ij} \mathcal{L}_{ij})_{;k} = \sum_{i,j} g^{ij} \mathcal{L}_{ij;k}, \quad \text{again by Ricci's Lemma.} \end{aligned}$$

This tells us that

$$(3c) \quad \sum_j C_{jj;k} = 0 \implies \sum_k \left(\sum_j C_{jj;k} \right) \theta^k(X) = 0$$

[which can also be obtained, with some pain, directly from the apolarity conditions $\sum_j c_{jjk} = 0$]. Finally, we note that

$$\begin{aligned} (3d) \quad - \sum_i \left(\sum_j C_{ij;j} \right) \theta^i(X) &= - \sum_j (\nabla_{X_j} \mathcal{J})(X, X_j) \\ &= - \text{trace}(Z \mapsto (\nabla_Z \mathcal{J})(X)), \end{aligned}$$

by the very same argument that was used on page 126. Now the equation

$$(3a) + (3b) = X(J - \mathcal{K}) + (3c) + (3d)$$

yields

The Special Affine Codazzi-Mainardi Equations:

$$X(J - \mathcal{K}) = \text{trace}(Z \mapsto (\nabla_Z \tilde{\mathcal{J}})(X)) - \mathcal{J} * \mathfrak{I}(X).$$

In terms of a map $f: U \rightarrow \mathbb{R}^3$ we have

$$(J - \mathcal{K})_j = \sum_{\mu, i} g^{\mu i} \mathcal{L}_{\mu j; i} - \sum_{i, \mu, \rho, \sigma} g^{\mu i} g^{\rho \sigma} \mathcal{L}_{\sigma i} \ell_{\mu \rho j}.$$

Finally, we are ready to state

21. FUNDAMENTAL THEOREM OF SPECIAL AFFINE SURFACE THEORY (RADON; 1918).

(1) Let $M, \bar{M} \subset \mathbb{R}^3$ be two connected surfaces in \mathbb{R}^3 , both consisting entirely of elliptic points or both consisting entirely of hyperbolic points; in the former case give both surfaces the usual orientation, and in the latter case, suppose that each surface is also oriented. Let $\nu: M \rightarrow \mathbb{R}^3$ and $\bar{\nu}: \bar{M} \rightarrow \mathbb{R}^3$ be the affine normal vector fields (determined by the orientations), and let $\mathfrak{I}, \mathfrak{I}$ and $\bar{\mathfrak{I}}, \bar{\mathfrak{I}}$ be the first and second affine fundamental forms for M and \bar{M} , respectively. Let $\phi: M \rightarrow \bar{M}$ be an orientation preserving diffeomorphism such that

$$\phi^* \bar{\mathfrak{I}} = \mathfrak{I} \quad \text{and} \quad \phi^* \bar{\mathfrak{I}} = \mathfrak{I}.$$

Then there is a special linear affine motion $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that $\phi = A|_M$ and $A_* \nu = \bar{\nu}$.

(2) Let M be an oriented 2-manifold with a (not necessarily positive definite) metric $\langle \cdot, \cdot \rangle$ having covariant derivative ∇ , curvature tensor \mathcal{R} , and curvature \mathcal{K} . Let \mathcal{S} be a symmetric tensor on M of order 3. Define

$$J = \frac{1}{2} \langle \mathcal{S}, \mathcal{S} \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product on tri-linear maps determined by the inner product $\langle \cdot, \cdot \rangle$, and define $\mathcal{J}, \mathcal{J}, \tilde{\mathcal{J}}$ by

$$\begin{aligned} \langle \mathcal{J}(X, Y), Z \rangle &= \mathcal{S}(X, Y, Z) \\ \mathcal{J}(X, Y) &= \text{trace}(Z \mapsto (\nabla_Z \mathcal{J})(X, Y)) \\ \langle \tilde{\mathcal{J}}(X), Y \rangle &= \mathcal{J}(X, Y). \end{aligned}$$

Suppose that \mathcal{S} satisfies

(1) The Apolarity Conditions:

$$\text{trace}(X \mapsto \mathcal{A}(X, Y)) = 0$$

(2) The Special Affine Codazzi-Mainardi Equations:

$$X(J - \mathcal{K}) = \text{trace}(Z \mapsto (\nabla_Z \tilde{\mathcal{S}})(X)) - \mathcal{S} * \mathcal{S}(X).$$

Then for any point $p \in M$ there is a neighborhood U of p and an immersion $f: U \rightarrow \mathbb{R}^3$ such that

$$\begin{aligned} \langle\langle \cdot, \cdot \rangle\rangle &= f^* \mathbf{I} \\ \mathcal{S} &= f^* \mathbf{II}, \end{aligned}$$

where \mathbf{I} and \mathbf{II} are the affine first and second fundamental forms on $f(U)$ determined by the orientation $f(U)$ gets from the orientation on $U \subset M$.

This can be proved in the same way that we proved Theorem 3, using the classical integrability theorem (I.6-1). Or the Frobenius form of the integrability conditions can be used (see the treatment for ordinary surface theory in Chapter 7). One can also reduce the theorem to Theorems I.10-17 and I.10-18; our integrability conditions reduce to the equations of structure of $\text{SL}(3, \mathbb{R})$.

PROBLEMS

1. Consider the Codazzi-Mainardi equations in Corollary 1-12, but write $(\nabla_X \Pi)(Y, Z) = X(\Pi(Y, Z)) - \Pi(\nabla_X Y, Z) - \Pi(Y, \nabla_X Z)$, and similarly for $(\nabla_Y \Pi)(X, Z)$. Choose $X = f_1$, $Y = f_2$, and then $Z = f_1$ or f_2 , to obtain equations (B') on page 56.

2. From equation (**) on page 53 show that

$$N_{ij} = - \sum_h \left(l_{i,j}^h + \sum_\rho l_i^\rho \Gamma_{\rho j}^h \right) f_h - \sum_{h,\rho} g^{h\rho} l_{\rho i} l_{hj} N.$$

Conclude that the equation $N_{ij} = N_{ji}$ is equivalent to

$$l_{i,j}^h + \sum_\rho l_i^\rho \Gamma_{\rho j}^h = l_{j,i}^h + \sum_\rho l_j^\rho \Gamma_{\rho i}^h.$$

Write $l_i^h = \sum_k g^{hk} l_{ki}$, and similarly for l_j^h , and expand. Multiply by $\sum_h g_{\tau h}$, and then use $\sum_h g_{\tau h} g^{hk}{}_{,j} = - \sum_h g_{\tau h,j} g^{hk}$. Show that the resulting equation is equivalent to the Codazzi-Mainardi equations, by making use of the identity

$$g_{ik,j} = [ij, k] + [jk, i].$$

3. Use the method of Problem 1-5 to prove that we can take U to be all of M in Corollary 5 when M is simply-connected.

4. (a) Find a continuous map $f: \mathbb{R} \rightarrow S^1$ which is onto and locally one-one, but not a covering map. *Hint:* Take part of the universal covering space of S^1 .
 (b) Let $f: X \rightarrow Y$ be a continuous map which is onto and locally a homeomorphism, and let X be compact. Then for every $y \in Y$, the set $f^{-1}(y) \subset X$ is finite, say $f^{-1}(y) = \{x_1, \dots, x_k\}$. Choose disjoint open sets $U_i \ni x_i$ such that f is a homeomorphism on each U_i , and let $U = \bigcap_i f(U_i)$. Using compactness of X , show that there is a compact neighborhood $K \subset U$ of y such that $f^{-1}(K) \subset \bigcup_i U_i$. Conclude that f is a covering map.

5. Let X_1, \dots, X_n be a moving frame on a manifold with a connection ∇ . Let

$$A = \sum a_{i_1 \dots i_k} \theta^{i_1} \otimes \dots \otimes \theta^{i_k}$$

be a tensor field of type $\binom{k}{0}$, and let

$$\nabla A = \sum a_{i_1 \dots i_k; l} \theta^{i_1} \otimes \dots \otimes \theta^{i_k} \otimes \theta^l.$$

Use Problem 1-1 to check that

$$\sum_l a_{i_1 \dots i_k; l} \theta^l = da_{i_1 \dots i_k} - \sum_\rho a_{\rho i_2 \dots i_k} \omega_{i_1}^\rho - \dots - \sum_\rho a_{i_1 \dots i_{k-1} \rho} \omega_{i_k}^\rho.$$

CHAPTER 3

A COMPENDIUM OF SURFACES

In the following chapters it will often be quite useful to have a detailed knowledge of the classical surfaces. In this chapter we will list all the important ones systematically, together with many of their properties; other properties will be mentioned in later chapters, when we have the theorems necessary to derive them. A brief survey of the initial pages may prove quite discouraging, but the reader can be assured that after the basic formulas have been derived once and for all, the reading becomes a lot more pleasant—there are pretty pictures to draw, and interesting points to be made.

Usually we will represent a surface locally as the image of an immersion $f: U \rightarrow \mathbb{R}^3$. We collect here the formulas from the previous chapters, with a few additions which are used for actual calculations.

$$\begin{array}{l}
 \text{(A)} \quad \begin{array}{ccc}
 E = \langle f_1, f_1 \rangle & F = \langle f_1, f_2 \rangle & G = \langle f_2, f_2 \rangle \\
 \\
 N = \frac{f_1 \times f_2}{|f_1 \times f_2|} = \frac{f_1 \times f_2}{\sqrt{EG - F^2}} \\
 \\
 \begin{array}{ccc}
 l = \langle -N_1, f_1 \rangle & m = \langle -N_1, f_2 \rangle & n = \langle -N_2, f_2 \rangle \\
 = \langle N, f_{11} \rangle & = \langle N, f_{12} \rangle & = \langle N, f_{22} \rangle \\
 \\
 \det \begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix} & \det \begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix} & \det \begin{pmatrix} f_{22} \\ f_1 \\ f_2 \end{pmatrix} \\
 = \frac{\det \begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix}}{\sqrt{EG - F^2}} & = \frac{\det \begin{pmatrix} f_{12} \\ f_1 \\ f_2 \end{pmatrix}}{\sqrt{EG - F^2}} & = \frac{\det \begin{pmatrix} f_{22} \\ f_1 \\ f_2 \end{pmatrix}}{\sqrt{EG - F^2}}
 \end{array}
 \end{array}
 \end{array}$$

(Here we have made a specific choice of N , which will influence the sign of various quantities to be computed later on.) We also recall that

$$\begin{aligned}
 \left(\begin{array}{l} \text{matrix of } -dv: M_p \rightarrow M_p \\ \text{with respect to } (f_1)_p, (f_2)_p \end{array} \right) &= (g_{ij})^{-1}(l_{ij}) \\
 &= \frac{1}{EG - F^2} \cdot \begin{pmatrix} G & -F \\ -F & E \end{pmatrix} \begin{pmatrix} l & m \\ m & n \end{pmatrix}
 \end{aligned}$$

[f_i, g_{ij}, l_{ij} evaluated at (s, t) ; where $p = f(s, t)$].

The principal curvatures k_1, k_2 are the eigenvalues of this matrix, and the Gaussian curvature K is $k_1 \cdot k_2$, while the mean curvature H is $(k_1 + k_2)/2$. So K and H are the determinant and half the trace, respectively, of this matrix. This gives us

$$(B) \quad \boxed{\begin{aligned} K &= \frac{ln - m^2}{EG - F^2} \\ H &= \frac{En - 2Fm + Gl}{2(EG - F^2)}. \end{aligned}}$$

(The sign of H depends on the choice of N , but the sign of K does not.)

In equation (B), the left sides H and K must be evaluated at $p = f(s, t)$ when the right sides are evaluated at (s, t) ; similar conventions will be used in the remaining equations. We remind the reader that it is also possible, in principle at least, to compute K directly from E, F, G . Since k_1, k_2 are the roots of the equation $\lambda^2 - 2H\lambda + K = 0$, we have

$$(C) \quad \boxed{k_1, k_2 = H \pm \sqrt{H^2 - K}.}$$

(The signs of k_1, k_2 both change when N is changed.)

It is, as usual, rather more difficult to find the principal directions, that is, the *eigenvectors* of $-dv$. We leave it to the reader (Problem 1) to show that

$$(D) \quad \boxed{\begin{aligned} &a_1 f_1 + a_2 f_2 \text{ is a principal vector if and only if} \\ &\det \begin{pmatrix} a_2^2 & -a_1 a_2 & a_1^2 \\ E & F & G \\ l & m & n \end{pmatrix} = 0. \end{aligned}}$$

We have already pointed out that at an umbilic point we have the necessary and sufficient condition

$$(E) \quad \boxed{l = kE, \quad m = kF, \quad n = kG \quad \text{at an umbilic point,} \\ \text{with } k_1 = k_2 = k.}$$

(The sign of k depends on N .)

This condition can also be derived from (D), since the determinant must be 0 for all choices of a_1 and a_2 . It is also clear that

$$(F) \quad \boxed{\begin{aligned} &a_1 f_1 + a_2 f_2 \text{ is an asymptotic vector if and only if} \\ &la_1^2 + 2ma_1 a_2 + na_2^2 = 0. \end{aligned}}$$

Finally, it is easy to see that

(G) If $F = m = 0$ at a point, then f_1 and f_2 are principal vectors there, and

$$-dv(f_1) = \frac{l}{E} f_1, \quad -dv(f_2) = \frac{n}{G} f_2.$$

When our surface is actually the graph of a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$, so that we can choose

$$f(s, t) = (s, t, h(s, t))$$

$$f_1 = (1, 0, h_1(s, t))$$

$$f_2 = (0, 1, h_2(s, t)) \quad f_{ij} = (0, 0, h_{ij}),$$

we obtain the following formulas:

(A')

$$E = 1 + h_1^2 \quad F = h_1 h_2 \quad G = 1 + h_2^2$$

$$N = \frac{(-h_1, -h_2, 1)}{\sqrt{1 + h_1^2 + h_2^2}}$$

$$l = \frac{h_{11}}{\sqrt{1 + h_1^2 + h_2^2}} \quad m = \frac{h_{12}}{\sqrt{1 + h_1^2 + h_2^2}} \quad n = \frac{h_{22}}{\sqrt{1 + h_1^2 + h_2^2}}.$$

(B')

$$K = \frac{h_{11}h_{22} - h_{12}^2}{[1 + h_1^2 + h_2^2]^2}$$

$$H = \frac{(1 + h_1^2)h_{22} + (1 + h_2^2)h_{11} - 2h_1h_2h_{12}}{2[1 + h_1^2 + h_2^2]^{3/2}}.$$

(C')

$$k_1, k_2 = H \pm \sqrt{H^2 - K}.$$

(D')

$$a_1 f_1 + a_2 f_2 = (a_1, a_2, a_1 h_1 + a_2 h_2) \text{ is a principal vector if and only if}$$

$$\det \begin{pmatrix} a_2^2 & -a_1 a_2 & a_1^2 \\ 1 + h_1^2 & h_1 h_2 & 1 + h_2^2 \\ h_{11} & h_{12} & h_{22} \end{pmatrix} = 0.$$

(E')

$$h_{11} = k(1 + h_1^2), \quad h_{12} = kh_1h_2, \quad h_{22} = k(1 + h_2^2)$$

at an umbilic point, with $k_1 = k_2 = k\sqrt{1 + h_1^2 + h_2^2}$.

(F')

$$a_1f_1 + a_2f_2 = (a_1, a_2, a_1h_1 + a_2h_2) \text{ is an}$$

asymptotic vector if and only if

$$h_{11}a_1^2 + 2h_{12}a_1a_2 + h_{22}a_2^2 = 0.$$

It is also quite useful to be able to compute K and H for surfaces

$$M = \{p \in \mathbb{R}^2 : W(p) = 0\},$$

where $W: \mathbb{R}^3 \rightarrow \mathbb{R}$. Recall (pg. II.113) that we can choose

$$v = \frac{Z}{|Z|} \quad \text{for } Z = (W_1, W_2, W_3).$$

If $X = (a_1, a_2, a_3)_p$, then

$$\begin{aligned} -dv(X) &= -\nabla_X \frac{Z}{|Z|} = -\frac{1}{|Z|} \nabla_X Z - X \left(\frac{1}{|Z|} \right) Z \\ &= -\frac{1}{|Z|} \left(\sum_i a_i W_{1i}, \sum_i a_i W_{2i}, \sum_i a_i W_{3i} \right) - \underbrace{X \left(\frac{1}{|Z|} \right) Z}_{\text{normal to } M}, \end{aligned}$$

so

$$\langle -dv(X), X \rangle = -\frac{1}{|Z|} \sum_{i,j} a_i a_j W_{ij} \quad [W_{ij} \text{ evaluated at } p].$$

This means that the principal curvatures k_1, k_2 at p are $-1/|Z|$ times the maximum and minimum of

$$\sum_{i,j} a_i a_j W_{ij} \quad \text{on } S = \{(a_1, a_2, a_3) : \sum_i a_i^2 = 1 \text{ and } \sum_i a_i W_i = 0\}$$

$[W_i, W_{ij} \text{ evaluated at } p].$

Using Lagrangian multipliers (see Problem 3, if you have forgotten them), these extrema occur at $(a_1, a_2, a_3) \in S$ if and only if there are λ, μ such that

$$D_j \left(\sum_i a_i a_k W_{ik} \right) = \lambda D_j \left(\sum_i a_i^2 \right) + \mu D_j \left(\sum_i a_i W_i \right) \quad \text{for all } j.$$

Using $W_{ik} = W_{ki}$, we find that

$$\begin{aligned} \sum_i a_i W_{ij} &= \lambda a_j + \frac{\mu}{2} W_j \\ \sum_i a_i^2 &= 1, \quad \sum_i a_i W_i = 0 \end{aligned} \quad \text{for some } \lambda, \mu.$$

Since we then have $\sum_{i,j} a_i a_j W_{ij} = \lambda$, this shows that the desired maximum and minimum values of $\sum_{i,j} a_i a_j W_{ij}$ are precisely the numbers λ for which we have

$$\begin{cases} \sum_i a_i W_{ij} = \lambda a_j + \frac{\mu}{2} W_j \\ \sum_i a_i W_i = 0 \end{cases} \quad \text{for some } (a_1, a_2, a_3) \neq 0, \text{ and some } \mu.$$

We can also write this as

$$(*) \quad \begin{cases} (W_{ij}) \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \frac{\mu}{2} \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \\ \sum_i a_i W_i = 0 \end{cases} \quad \text{for some } (a_1, a_2, a_3) \neq 0, \text{ and some } \mu.$$

Now, since

$$\begin{pmatrix} \boxed{(W_{ij}) - \lambda I} & \begin{matrix} W_1 \\ W_2 \\ W_3 \end{matrix} \\ \begin{matrix} W_1 & W_2 & W_3 \end{matrix} & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ t \end{pmatrix} = \begin{pmatrix} (W_{ij}) \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} - \lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + t \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} \\ a_1 W_1 + a_2 W_2 + a_3 W_3 \end{pmatrix},$$

we see that (*) holds precisely when there exists $(a_1, a_2, a_3, t) \neq 0$ so that the right side of the above equation is the 0 column vector. Thus the desired λ 's are those for which the left-hand 4×4 matrix has determinant 0:

$$(C'') \quad \boxed{\begin{aligned} k_i &= -\frac{1}{\sqrt{W_1^2 + W_2^2 + W_3^2}} \lambda_i, \\ \text{where } \lambda_i &\text{ are the roots of the quadratic equation} \\ (**) \quad \det \begin{pmatrix} \boxed{(W_{ij}) - \lambda I} & \begin{matrix} W_1 \\ W_2 \\ W_3 \end{matrix} \\ \begin{matrix} W_1 & W_2 & W_3 \end{matrix} & 0 \end{pmatrix} &= 0. \end{aligned}}$$

Consequently,

$$(B'') \quad \begin{aligned} K &= \frac{1}{W_1^2 + W_2^2 + W_3^2} \cdot \frac{\text{constant term in } (**)}{\text{coefficient of } \lambda^2 \text{ in } (**)} \\ H &= -\frac{1}{2\sqrt{W_1^2 + W_2^2 + W_3^2}} \cdot \frac{\text{coefficient of } \lambda \text{ in } (**)}{\text{coefficient of } \lambda^2 \text{ in } (**)}. \end{aligned}$$

$$(C'') \quad \begin{aligned} k_1, k_2 &= \\ H \pm \sqrt{H^2 - K}. \end{aligned}$$

In particular, to anticipate the case of greatest interest for us, we find:

$$(***) \quad \text{If } (W_{ij}) = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \text{ then}$$

$$K = \frac{1}{(W_1^2 + W_2^2 + W_3^2)^2} [W_1^2 \lambda_2 \lambda_3 + W_2^2 \lambda_1 \lambda_3 + W_3^2 \lambda_1 \lambda_2]$$

$$H = \frac{1}{2(W_1^2 + W_2^2 + W_3^2)^{3/2}} [W_1^2(\lambda_2 + \lambda_3) + W_2^2(\lambda_1 + \lambda_3) + W_3^2(\lambda_1 + \lambda_2)].$$

[In all the formulas given so far, the sign of k_1, k_2 and of H depends on the choice of N as the normalized vector (W_1, W_2, W_3) .]

Since X is a principal vector if and only if $\langle d\nu(X) \times X, \nu \rangle = 0$, we also see that

$$(D'') \quad \begin{aligned} X = (a_1, a_2, a_3)_p \text{ is a principal vector if and only if} \\ \det \begin{pmatrix} W_1 & \sum_i a_i W_{1i} & a_1 \\ W_2 & \sum_i a_i W_{2i} & a_2 \\ W_3 & \sum_i a_i W_{3i} & a_3 \end{pmatrix} = 0 \quad \text{and} \quad \sum_i a_i W_i = 0. \end{aligned}$$

There does not seem to be any especially simple condition for an umbilic, but we do know that p is an umbilic if and only if the determinant in (D'') is 0

for all (a_1, a_2, a_3) with $\sum_i a_i W_i = 0$. Finally, we have

$$(F'') \quad \boxed{X = (a_1, a_2, a_3)_p \text{ is an asymptotic vector if and only if} \\ \sum_i a_i a_j W_{ij} = 0 \quad \text{and} \quad \sum_i a_i W_i = 0.}$$

We are now ready to begin our systematic list of surfaces. They come under five headings.

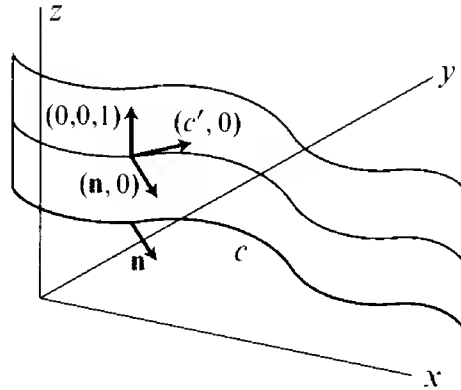
I. THE CLASSICAL FLAT SURFACES

1. Plane

For any plane, the normal map v is constant, so $dv = 0$. Thus all points are umbilics with $k_1 = k_2 = 0$, and $K = H = 0$. All vectors are also asymptotic. The plane is actually a special case of

2. Generalized Cylinder

Here our surface is $M = \{(x, y, z) : (x, y) = c(s) \text{ for some } s\}$, where c is an immersed curve in \mathbb{R}^2 , which we assume parameterized by arclength.



The normal map v is always parallel to the (x, y) plane. One principal direction at any point is $(0, 0, 1)$, with $k_1 = 0$. The other principal direction at $(c(s), z)$ is $(c', 0) = (\mathbf{n}, 0)$, with $k_2 = \kappa(s)$, the curvature of c at s (here we choose the normal map v to be $(\mathbf{n}, 0)$, where the normal \mathbf{n} for c is picked as on pg. II.6; recall that $\mathbf{n}'(s) = -\kappa(s) \cdot c'(s)$, while k_2 is an eigenvalue of $-dv$). Hence $K = 0$ and $H = \frac{1}{2}\kappa(s)$.

The only asymptotic direction is $(0, 0, 1)$, unless $k(s) = 0$, in which case all directions are asymptotic.

3. Generalized Cone

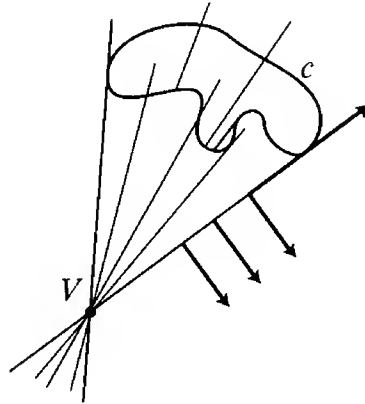
Our surface is parameterized by

$$f(s, t) = V + t[c(s) - V],$$

where $V \in \mathbb{R}^3$ is the *vertex*, and c is an immersed curve in \mathbb{R}^3 . For f to be an immersion, the vectors

$$f_1 = tc' \quad \text{and} \quad f_2 = c - V$$

must be linearly independent, so we must have $t \neq 0$ (V cannot be in the surface) and $c'(s)$ linearly independent of $c(s) - V$.



Since the tangent space at $f(s, t)$ is spanned by $c'(s)$ and $c(s) - V$, the normal map is constant along the straight lines obtained by keeping s fixed. Consequently, the vectors $f_2(s, t) = c(s) - V$ are principal vectors, with $k_1 = 0$. Therefore $K = 0$. Once again, these vectors are also asymptotic, and there are no others, except at points where $H = k_2/2$ happens to be 0, in which case all vectors are asymptotic.

4. Tangent Developable

This surface consists of the tangents to a curve c in \mathbb{R}^3 (parameterized by arclength as usual). It can therefore be parameterized by

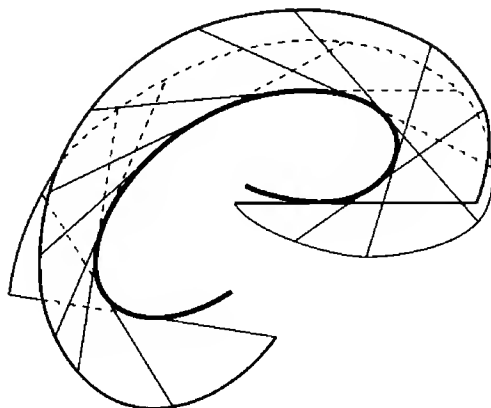
$$f(s, t) = c(s) + tc'(s).$$

We have

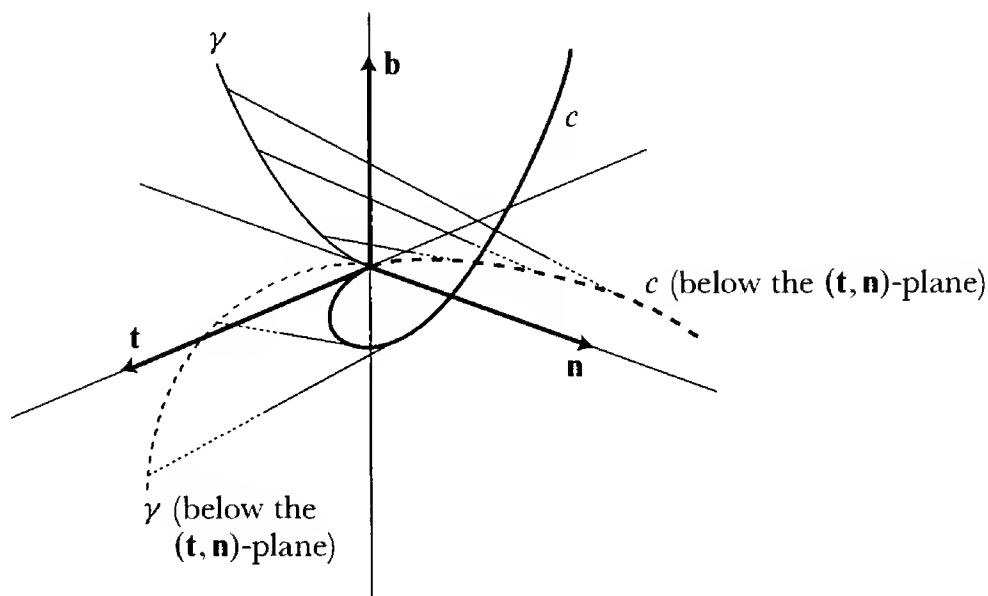
$$\begin{aligned} f_1 &= c' + tc'' = c' + t\kappa \mathbf{n}, & \text{where } \mathbf{n} \text{ is the normal vector} \\ & & \text{of } c, \text{ and } \kappa \text{ is the curvature} \\ f_2 &= c'; \end{aligned}$$

so f is regular if $t \neq 0$ and $\kappa \neq 0$. Thus this parameterization does not allow

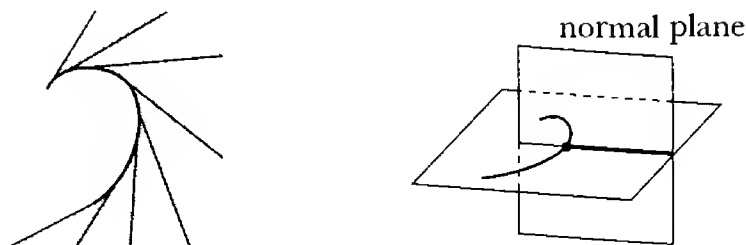
the curve itself to be part of the surface. In fact, the surface consists of two sheets which meet along the curve in a sharp edge (the **edge of regression** or



cuspidal edge); at any point of the curve, the normal plane intersects the surface in a curve γ with a cusp. This is shown below for our “standard curve” of pp. II.30, 31.



Actually, the assertion we have just made requires some careful interpretation, as can be seen by considering the case where c lies in a plane. The two sheets of the tangent developable are then the same portion of this plane, and their intersection with a normal plane is just a ray. In general, we can analyze the



intersection of the tangent developable and a normal plane to c as follows. For convenience, we consider the point $s = 0$, and assume $c(0) = 0$ and that $\mathbf{t}, \mathbf{n}, \mathbf{b}$ lie along the three coordinate axes. For small $s \neq 0$, the vector $c'(s)$, being close to $\mathbf{t} = c'(0)$, does not lie in the (\mathbf{n}, \mathbf{b}) -plane; moreover, $c(s)$ also does not lie in the (\mathbf{n}, \mathbf{b}) -plane. Now we can write (using the Serret-Frenet formulas)

$$\begin{aligned} c(s) &= c(0) + sc'(0) + \frac{s^2}{2}c''(0) + \frac{s^3}{6}c'''(0) + o(s^3) \\ &= (0, 0, 0) + s(1, 0, 0) + \frac{s^2}{2}(0, \kappa, 0) + \frac{s^3}{6}(-\kappa^2, \kappa', \kappa\tau) + o(s^3); \end{aligned}$$

here $\kappa = \kappa(0)$ and $\tau = \tau(0)$, and $o(s^3)$ is a function with the property that $o(s^3)/s^3 \rightarrow 0$ as $s \rightarrow 0$. Similarly, we have

$$\begin{aligned} c'(s) &= c'(0) + sc''(0) + \frac{s^2}{2}c'''(0) + o(s^2) \\ &= (1, 0, 0) + s(0, \kappa, 0) + \frac{s^2}{2}(-\kappa^2, \kappa', \kappa\tau) + o(s^2). \end{aligned}$$

Combining, we have

$$\begin{aligned} (*) \quad f(s, t) &= c(s) + tc'(s) \\ &= (0, 0, 0) + s(1, 0, 0) + \frac{s^2}{2}(0, \kappa, 0) + \frac{s^3}{6}(-\kappa^2, \kappa', \kappa\tau) + o(s^3) \\ &\quad + t \left[(1, 0, 0) + s(0, \kappa, 0) + \frac{s^2}{2}(-\kappa^2, \kappa', \kappa\tau) + o(s^2) \right]. \end{aligned}$$

For each s there is $t(s)$ for which $f(s, t(s))$ lies in the (\mathbf{n}, \mathbf{b}) -plane, which means that the first component of $f(s, t(s))$ equals zero:

$$s - \frac{\kappa^2}{6}s^3 + o(s^3) + t(s) \left[1 - \frac{\kappa^2}{2}s^2 + o(s^2) \right] = 0.$$

Dividing through, and giving just a little thought to the meaning of what we are doing, we find that

$$t(s) = -s - \frac{\kappa^2}{3}s^2 + o(s^3).$$

[Here, and below, $o(s^3)$ represents some *new* function with the property that $o(s^3)/s^3 \rightarrow 0$ as $0 \rightarrow 0$.] Substituting back into (*), we find that

$$\begin{aligned} 2^{\text{nd}} \text{ component of } f(s, t(s)) &= \frac{\kappa}{2}s^2 + o(s^2) + t(s)\{\kappa s + o(s)\} \\ &= -\frac{\kappa}{2}s^2 + o(s^2), \end{aligned}$$

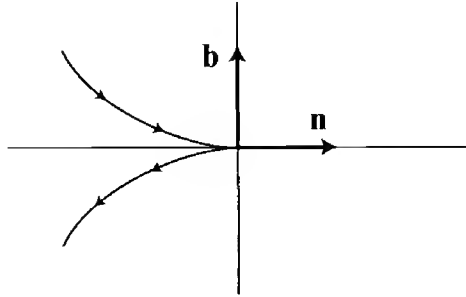
$$\begin{aligned} 3^{\text{rd}} \text{ component of } f(s, t(s)) &= \frac{\kappa\tau}{6}s^3 + o(s^3) + t(s)\left\{\frac{\kappa\tau}{2}s^2 + o(s^2)\right\} \\ &= -\frac{\kappa\tau}{3}s^3 + o(s^3). \end{aligned}$$

Thus, in the (\mathbf{n}, \mathbf{b}) -plane, or, in other words, in the (y, z) -plane, the intersection is described up to first order as the curve

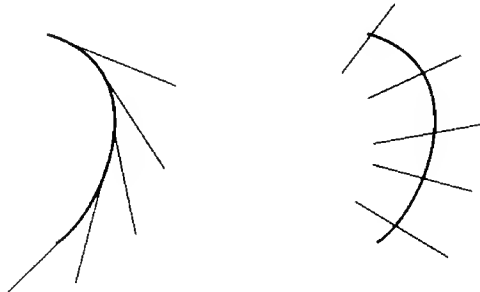
$$s \mapsto \left(-\frac{\kappa}{2}s^2, -\frac{\kappa\tau}{3}s^3\right) = (y(s), z(s));$$

the image is the graph of

$$z^2 = -\frac{8}{9} \frac{\tau^2}{\kappa} y^3.$$



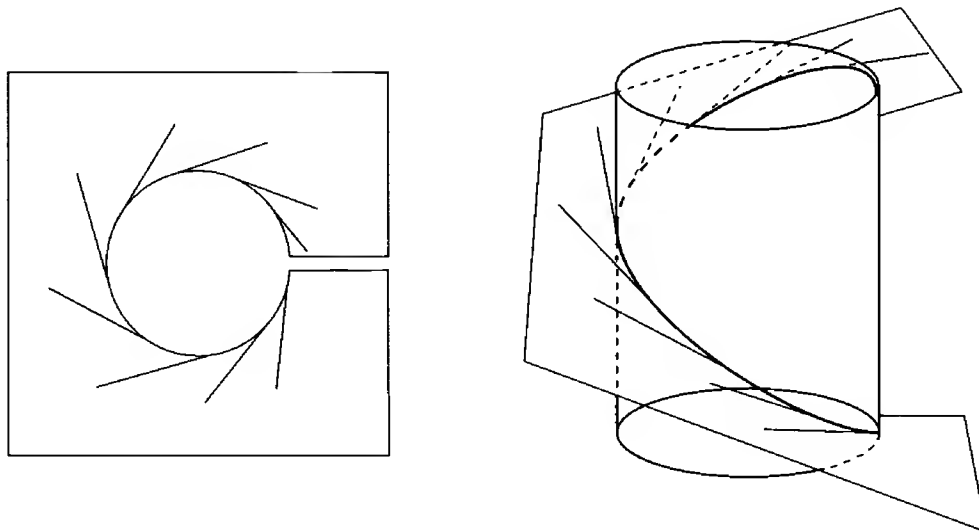
This analysis leads one to suspect that a *single* branch of the tangent developable can be extended so as to include the original curve, though, to be sure, a different parameterization is required. This is most obvious for a plane curve; one obtains an immediate extension if one uses lines perpendicular to the curve as t -parameter lines. Presumably in the general case we obtain a nice param-



eterization of $\{c(s) + tc'(s) : t \geq 0\}$ when we choose as one set of parameter lines the intersection of the surface with normal planes to the curve.

Since the tangent space at $f(s, t)$ is spanned by $c'(s)$ and $\mathbf{n}(s)$, it is the same along the straight lines obtained by keeping s fixed. So the normal map is constant along these lines, and the vectors $f_2(s, t) = c'(s)$ are principal vectors, with $k_1 = 0$. Once again, $K = 0$.

Since all the surfaces in our first category are flat, they are all locally isometric to the plane (that is the reason for the name “tangent developable”—a “development” of one surface on another is the very classical name for an isometry). The reader may easily construct isometries between the plane and generalized cylinders or cones. To map the tangent developable of c isometrically on the plane, we note that f_1, f_2 , and consequently E, F, G , depend only on the curvature κ of c , *not on its torsion*. So if c_1 is a plane curve with the same curvature as c , then our original surface is isometric to the tangent developable of c_1 , which is a subset of the plane. As an application of this fact, we note that the tangent developable of a helix (which has constant curvature) can be constructed from a piece of paper by cutting a circle out, and twisting the remaining portion around a cylinder.



All the surfaces in our first category are special cases of the surfaces in our second.

II. RULED SURFACES

These are the surfaces which can be parameterized as

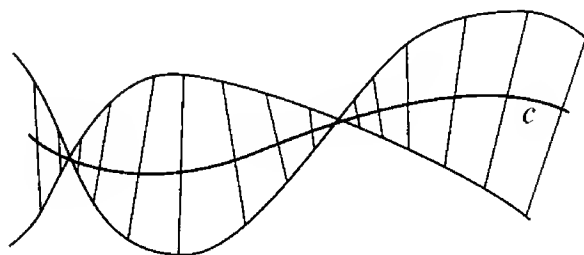
$$f(s, t) = c(s) + t\delta(s)$$

for two curves c, δ . Since

$$f_1 = c' + t\delta' \quad \text{and} \quad f_2 = \delta,$$

the map f is an immersion when δ and $c' + t\delta'$ are linearly independent. In

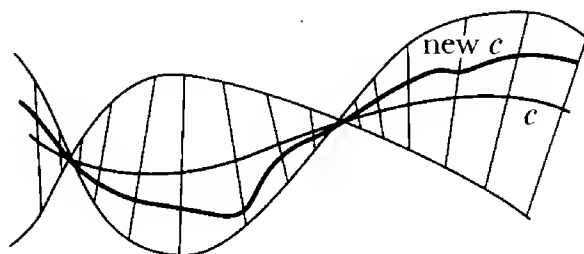
particular, if δ and c' are linearly independent, this certainly happens for sufficiently small t . For fixed s , we obtain straight lines [or segments] through $c(s)$; these straight lines are called the *rulings* of the surface. The ruled surfaces which are not generalized cones or cylinders, or tangent developables, are sometimes called *scrolls*.



A calculation shows (Problem 4) that

$$K = \frac{-m^2}{EG - F^2} = \frac{-\langle c', \delta \times \delta' \rangle^2}{|(c' + t\delta') \times \delta|^2}.$$

A more reasonable formula for K is obtained when we choose our parameterization more carefully. Note first that we might as well choose δ to be a curve with $|\delta(s)| = 1$ (and consequently $\langle \delta(s), \delta'(s) \rangle = 0$); it is then called the **directrix** of the surface. Next note that if c is replaced by any curve which intersects each ruling only once, and the directrix is kept the same, then we obtain the same



surface. Let us assume that we always have $\delta'(s) \neq 0$ ("the directions of the rulings are always changing"). Then the ruling L_s through $c(s)$ is not parallel to the ruling $L_{s+\varepsilon}$ for small ε , so there is a unique point $P(\varepsilon)$ on L_s closest to $L_{s+\varepsilon}$. One can show (Problem 5) that as $\varepsilon \rightarrow 0$, the point $P(\varepsilon)$ approaches the point

$$\sigma(s) = c(s) - \frac{\langle c'(s), \delta'(s) \rangle}{\langle \delta'(s), \delta'(s) \rangle} \cdot \delta(s) \quad \text{on } L_s.$$

We easily find from this that $\langle \sigma'(s), \delta'(s) \rangle = 0$. This is the only curve σ that can have this property: the point $\sigma(s)$ is simply the unique point on L_s where

the tangent plane of the surface contains a vector perpendicular to $\delta'(s)$. The curve σ is called the **striction curve** of the surface, and clearly depends only on the surface, not on the original parameterization. One further alteration eliminates all trace of the original parameterization—we might as well change s so that it is the arclength of δ . We thus have the “standard parameterization”

$$\begin{aligned} f(s, t) &= \sigma(s) + t\delta(s) \\ |\delta| &= |\delta'| = 1, \quad \langle \sigma', \delta' \rangle = 0 \\ [\delta, \sigma'] &\text{ linearly independent}. \end{aligned}$$

[The only thing that might go wrong in all this is that $\sigma(s)$ might be a point where the original, and hence the new, f is not an immersion. As a matter of fact (Problem 5), for the tangent developable of c , the striction curve is just c itself.] It now turns out (Problem 4) that

$$K = \frac{-p^2(s)}{(p^2(s) + t^2)^2} \quad \text{for } p = \langle \sigma'(s), \delta(s) \times \delta'(s) \rangle.$$

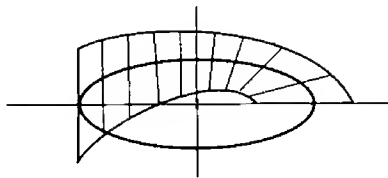
The function p is called the **distribution parameter**. Since p depends only on s , we see that $K \rightarrow 0$ as we go out along any ruling.

In addition to these general results, which we will put to use somewhat later, there are a few ruled surfaces of particular interest, two of which we will mention here, and two of which occur in our next category.

1. Möbius Strip

Our first example of a ruled surface is the “standard” Möbius strip, a slight modification of the one given on pg. I.10,

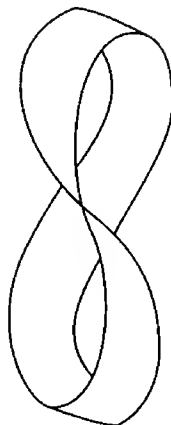
$$\begin{aligned} f(s, t) &= \left(\cos s + t \cos \frac{s}{2} \cos s, \sin s + t \cos \frac{s}{2} \sin s, t \sin \frac{s}{2} \right) \\ &= (\cos s, \sin s, 0) + t \left(\cos \frac{s}{2} \cos s, \cos \frac{s}{2} \sin s, \sin \frac{s}{2} \right) \quad (|t| < \frac{1}{2}). \end{aligned}$$



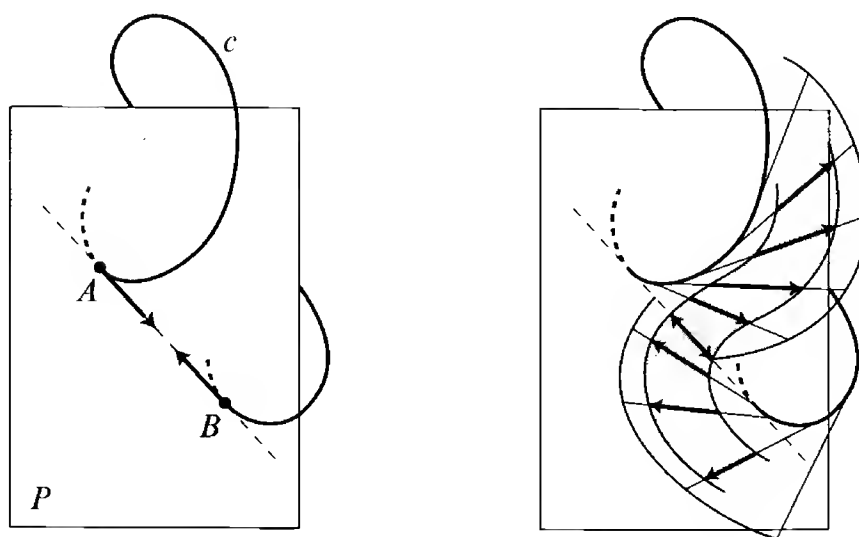
Computing directly from formulas (A) we find that

$$E = \frac{t^2}{4} + \left[1 + t \cos \left(\frac{s}{2} \right) \right]^2 \quad F = 0 \quad G = 1$$

(the values of F and G should be obvious!). Notice that this surface is *not* flat, so it is not the Möbius strip that one makes out of a strip of paper.



A C^∞ flat surface homeomorphic to the Möbius strip can be constructed as follows. We start with a curve c like the one shown in the first figure below.



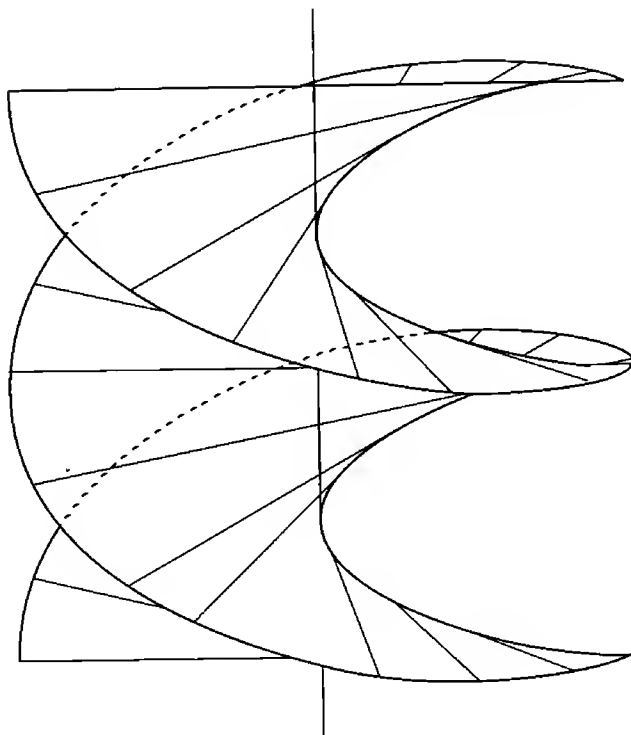
The tangent vectors at A and B are negatives of each other, and the plane P in which they lie is the osculating plane of c at these points, so that c'' lies in P at A and B . We then consider the portion of the tangent developable of c which is formed by the positive multiples of the tangent vectors. From this surface we can cut out a strip, as shown in the second part of the figure, which is homeomorphic to the Möbius strip. If we choose c so that all $c^{(k)}$, $k > 2$ vanish at A and B , then this surface will be C^∞ . Notice that the resulting picture is a “back-view” of the Möbius strip shown above.

In Chapter 5 we will mention a way of constructing an analytic flat Möbius strip.

2. (Right) Helicoid

This surface is generated by a line which moves along the z -axis in such a way that it remains parallel to the (x, y) -plane, and passes through the points of a circular helix (in other words, the surface is generated by a line perpendicular to the z -axis under a “screw” motion). It is thus given by

$$f(s, t) = (t \cos s, t \sin s, bs), \quad b \neq 0.$$



The lines $t = \text{constant}$ are helices (compare pg. II.32). Computing from (A) and (B), we find that

$$\begin{aligned} f_1(s, t) &= (-t \sin s, t \cos s, b) & f_2(s, t) &= (\cos s, \sin s, 0) \\ f_{11}(s, t) &= (-t \cos s, -t \sin s, 0) & f_{22}(s, t) &= (0, 0, 0) \\ f_{12}(s, t) &= (-\sin s, \cos s, 0) \\ E &= b^2 + t^2, \quad F = 0, \quad G = 1; & \sqrt{EG - F^2} &= \sqrt{b^2 + t^2} \\ l &= n = 0, \quad m = \frac{b}{\sqrt{b^2 + t^2}} \\ K &= -\frac{b^2}{(b^2 + t^2)^2}, \quad H = 0. \end{aligned}$$

The helices $t = \text{constant}$ intersect the rulings $s = \text{constant}$ in right angles ($F = 0$). They also point in the asymptotic directions ($l = n = 0$), so H must be 0.

III. QUADRIC SURFACES

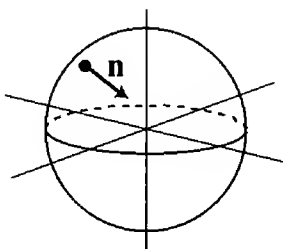
These are the surfaces of the form $W^{-1}(0)$, where

$$W(x_1, x_2, x_3) = \sum a_{ij} x_i x_j + \sum b_i x_i + c.$$

Standard arguments (Problem 6) show that, aside from trivial cases, they are all one of the following (up to rotations and translations); some of them are old friends of ours.

0. Sphere

This is a special case of the surfaces of the first group, but it surely deserves special mention. As we know, for the sphere of radius R , the normal map ν is just $-1/R$ times the identity (choosing the *inward* pointing normal, as on pg. II.52); every point is umbilic, the principal curvatures are $1/R$, and $K = 1/R^2$, $H = 1/2R$.

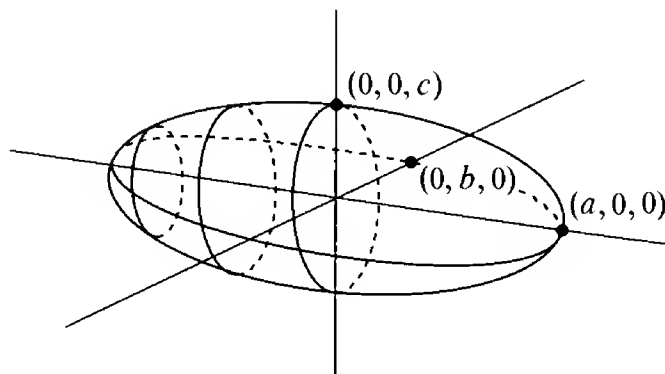


1. Ellipsoid

This surface has the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

The planes perpendicular to an axis intersect the surface in a family of similar ellipses.



Choosing

$$W(x, y, z) = \frac{1}{2} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 \right)$$

$$W_1(x, y, z) = \frac{x}{a^2} \quad W_2(x, y, z) = \frac{y}{b^2} \quad W_3(x, y, z) = \frac{z}{c^2}$$

$$(W_{ij}) = \begin{pmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & 1/c^2 \end{pmatrix}$$

and applying (***) on page 140, we have

$$K = \frac{1}{a^2 b^2 c^2} \cdot \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-2}.$$

Problem 7 casts this into a more useful form, which among other things allows us to see immediately that the maximum and minimum Gaussian curvatures occur at the expected places.

Using (D''), we find that (x, y, z) is an umbilic if and only if

$$\det \begin{pmatrix} x/a^2 & a_1/a^2 & a_1 \\ y/b^2 & a_2/b^2 & a_2 \\ z/c^2 & a_3/c^2 & a_3 \end{pmatrix} = 0 \text{ for all } a_1, a_2, a_3 \text{ with } \frac{a_1 x}{a^2} + \frac{a_2 y}{b^2} + \frac{a_3 z}{c^2} = 0.$$

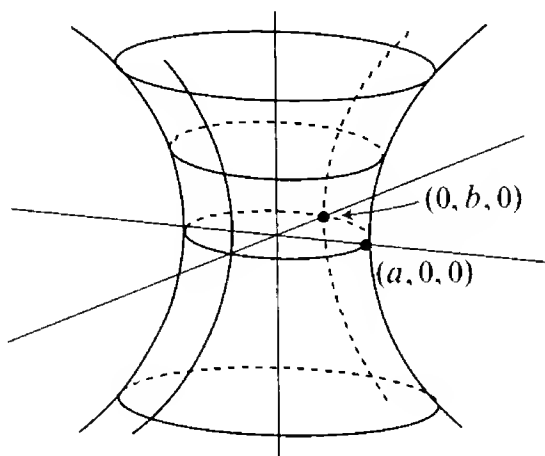
If $a > b > c > 0$, then there turn out to be exactly four umbilics on M (Problem 9). If, on the other hand, we are dealing with an ellipse of rotation, then there will be two whole circles of umbilics.

2. Elliptic Hyperboloid (of one sheet)

The equation now is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

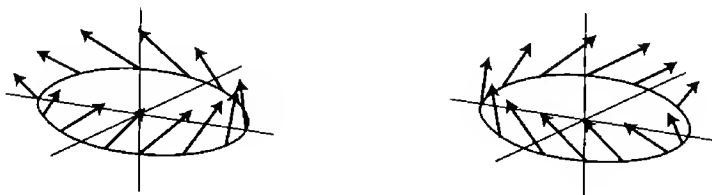
Planes perpendicular to the z -axis intersect the surface in similar ellipses, while planes perpendicular to the other axes intersect it in hyperbolas. When $a = b$, it may be obtained by revolving a hyperbola around the z -axis (hyperboloid of revolution).



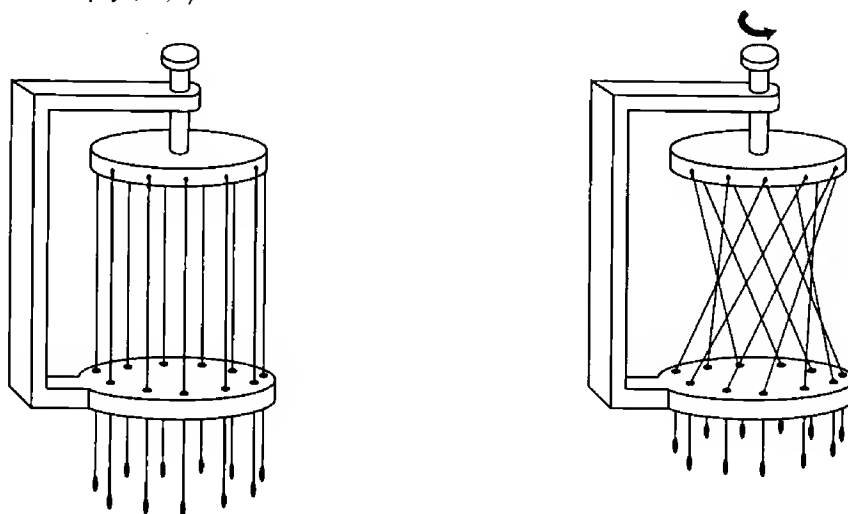
This surface is a *ruled surface*! In fact, it is doubly ruled: it may be parameterized as

$$\begin{aligned} f(s, t) &= (a \cos s, b \sin s, 0) + t(-a \sin s, b \cos s, c) \\ \text{or} \quad f(s, t) &= (a \cos s, b \sin s, 0) + t(a \sin s, -b \cos s, c). \end{aligned}$$

In each case the rulings pass through the ellipse $x^2/a^2 + y^2/b^2 = 1, z = 0$ and are perpendicular to the radius vector to that point. For a hyperboloid of



revolution, it is possible to demonstrate this ruling dramatically with apparatus like that pictured below. (The general elliptic hyperboloid must then also be ruled, since it is the image of an hyperboloid of revolution under a linear map $(x, y, z) \mapsto (\alpha x, \beta y, z)$.)



To compute K , we choose a W similar to that for the ellipsoid, and obtain

$$\begin{aligned} W_1 &= \frac{x}{a^2} & W_2 &= \frac{y}{b^2} & W_3 &= -\frac{z}{c^2} \\ (W_{ij}) &= \begin{pmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & 1/c^2 \end{pmatrix}. \end{aligned}$$

Then K turns out to be precisely

$$K = \frac{-1}{a^2 b^2 c^2} \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-2}.$$

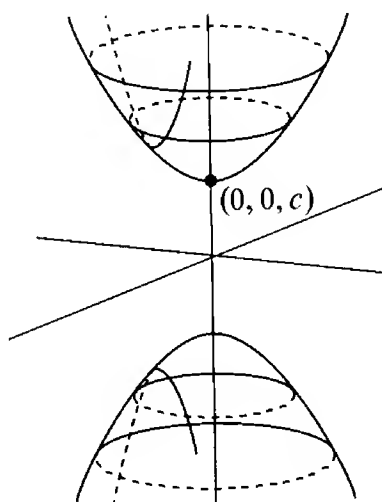
Since K is negative, there are, of course, no umbilics.

3. *Elliptic Hyperboloid (of two sheets)*

The equation is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1.$$

It is still true that planes perpendicular to the z -axis intersect the surface in ellipses (when they intersect it at all), while planes perpendicular to the other axes intersect it in hyperbolas. However, the surface looks quite different.



To compute K we choose a W which differs only by a constant from the W for the elliptic hyperboloid of one sheet. The computations are then precisely the same, except that the factor $x^2/a^2 + y^2/b^2 - z^2/c^2$ which appears is now equal to -1 . So we get

$$K = \frac{1}{a^2 b^2 c^2} \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-2}.$$

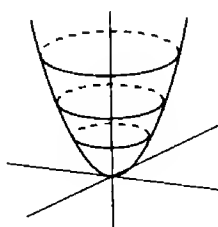
Again there are four umbilics (Problem 9).

4. *Elliptic Paraboloid*

The equation is simply

$$z = \frac{x^2}{a^2} + \frac{y^2}{b^2}.$$

Planes perpendicular to the z -axis intersect the surface in similar ellipses. Planes perpendicular to the other axes intersect it in parabolas. When $a = b$ we have a paraboloid of revolution.



Computations are especially simple:

$$\begin{aligned}
 W(x, y, z) &= \frac{1}{2} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} - z \right) \\
 W_1 &= \frac{x}{a^2} & W_2 &= \frac{y}{b^2} & W_3 &= -\frac{1}{2} \\
 (W_{ij}) &= \begin{pmatrix} 1/a^2 & 0 & 0 \\ 0 & 1/b^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
 K &= \frac{1}{4a^2b^2} \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{1}{4} \right)^{-2}.
 \end{aligned}$$

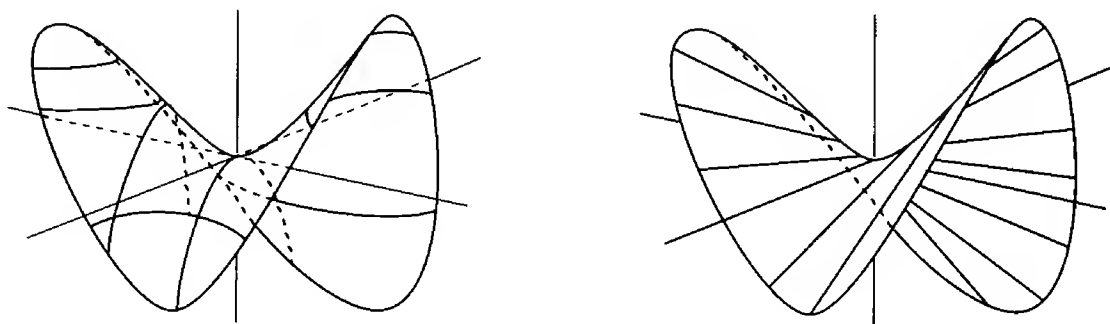
There are two umbilics (Problem 10).

5. *Hyperbolic Paraboloid*

Here the equation is

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}.$$

Planes perpendicular to the y -axis intersect the surface in parabolas, and planes perpendicular to the x -axis intersect the surface in parabolas pointing the other way. Planes perpendicular to the z -axis intersect the surface in hyperbolas pointing in one direction when the plane lies above the (x, y) -plane, and in the other direction when the plane lies below the (x, y) -plane; the (x, y) -plane itself intersects the surface in two intersecting straight lines.



This surface is also doubly ruled. It may be parameterized as

$$f(s, t) = (as, 0, s^2) + t(a, b, 2s) \quad \text{or} \quad f(s, t) = (as, 0, s^2) + t(a, -b, 2s).$$

[It is a classical result that all doubly ruled surfaces with $K < 0$ are quadratic. An elementary, somewhat unsatisfying proof is given in Problem 11; a nice proof can be given (Problem 4-16) by means of affine surface theory, which shouldn't

be too surprising, since the property of being doubly ruled is invariant under affine maps.]

For computations we choose

$$W(x, y, z) = \frac{1}{2} \left(\frac{x^2}{a^2} - \frac{y^2}{b^2} - z \right)$$

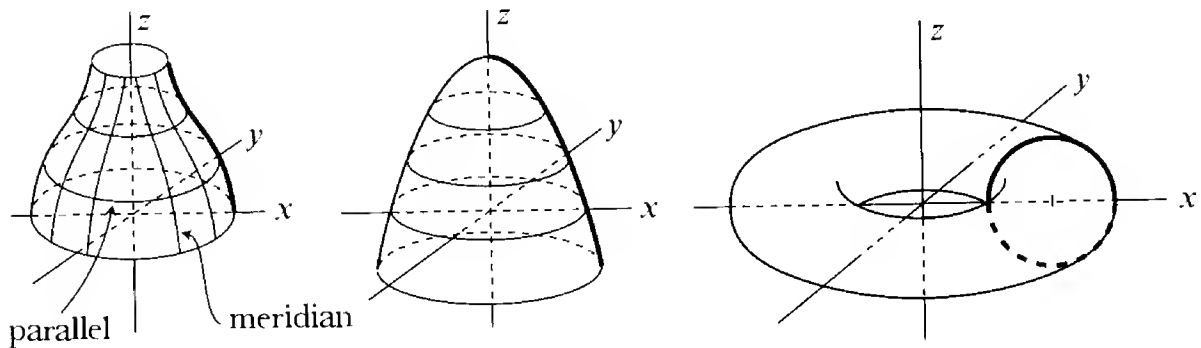
$$W_1 = \frac{x}{a^2} \quad W_2 = -\frac{y}{b^2} \quad W_3 = -\frac{1}{2}$$

$$(W_{ij}) = \begin{pmatrix} 1/a^2 & 0 & 0 \\ 0 & -1/b^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$K = \frac{-1}{4a^2b^2} \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{1}{4} \right)^{-2}.$$

IV. SURFACES OF REVOLUTION

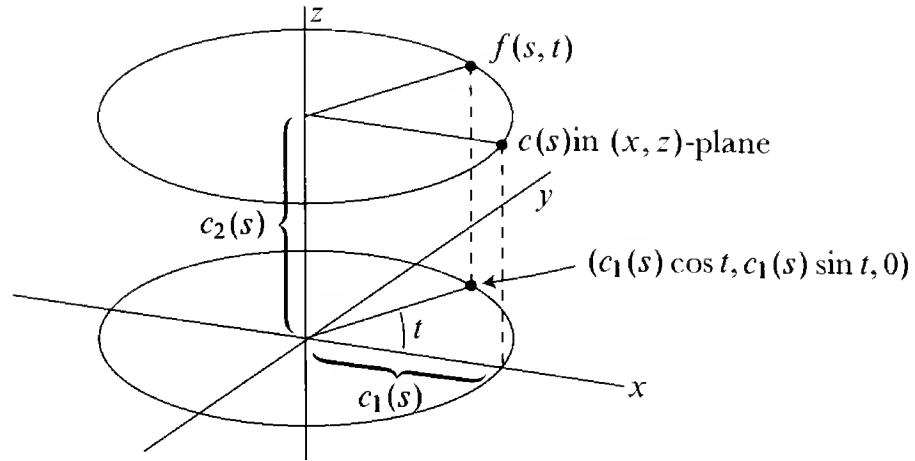
These are the surfaces obtained by starting with a curve (the **profile** curve), lying in the right half of the (x, z) -plane, and revolving it about the z -axis. If



the curve intersects the z -axis, it must do so at a right angle. As illustrated in the figure at the top of the next page, the surface is parameterized by

$$f(s, t) = (c_1(s) \cos t, c_1(s) \sin t, c_2(s)).$$

The curves $s = \text{constant}$ are called **parallels** and the curves $t = \text{constant}$ are called **meridians**.



By (A) we compute

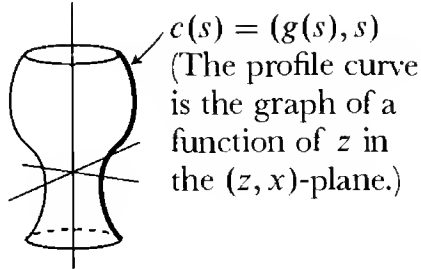
$$\begin{aligned}
 f_1 &= (c_1' \cos t, c_1' \sin t, c_2') & f_2 &= (-c_1 \sin t, c_1 \cos t, 0) \\
 f_{11} &= (c_1'' \cos t, c_1'' \sin t, c_2'') & f_{22} &= (-c_1 \cos t, -c_1 \sin t, 0) \\
 f_{12} &= (-c_1' \sin t, c_1' \cos t, 0) \\
 E &= (c_1')^2 + (c_2')^2 & F &= 0 & G &= c_1^2 \\
 \sqrt{EG - F^2} &= c_1 \sqrt{(c_1')^2 + (c_2')^2} \\
 l &= \frac{c_1' c_2'' - c_2' c_1''}{\sqrt{(c_1')^2 + (c_2')^2}} & m &= 0 & n &= \frac{c_1 c_2'}{\sqrt{(c_1')^2 + (c_2')^2}}
 \end{aligned}$$

Since $F = m = 0$, the tangent vectors of parallels and meridians point in the directions of the principal curvatures; we can use equations (G) to find directly that the principal curvatures are

$$\begin{aligned}
 k_{\text{meridian}} &= \frac{l}{E} = \frac{c_1' c_2'' - c_2' c_1''}{[(c_1')^2 + (c_2')^2]^{3/2}} \\
 k_{\text{parallel}} &= \frac{n}{G} = \frac{c_2'}{c_1 [(c_1')^2 + (c_2')^2]^{1/2}} \\
 K &= k_{\text{meridian}} \cdot k_{\text{parallel}} = \frac{c_2' (c_1' c_2'' - c_2' c_1'')}{c_1 [(c_1')^2 + (c_2')^2]^2} \\
 H &= \frac{1}{2} (k_{\text{meridian}} + k_{\text{parallel}}) .
 \end{aligned}$$

It will be useful for us to consider the special case where $c(s) = (g(s), s)$. Then we find that

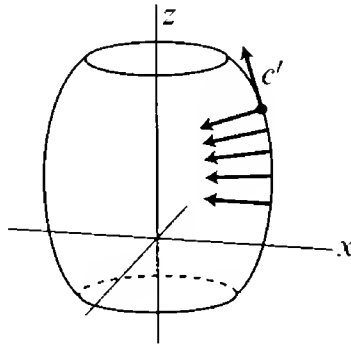
$$(3) \quad \begin{aligned} k_{\text{meridian}} &= \frac{-g''}{[1 + (g')^2]^{3/2}} \\ k_{\text{parallel}} &= \frac{1}{g[1 + (g')^2]^{1/2}} \\ K &= \frac{-g''}{g[1 + (g')^2]^2} \\ H &= \frac{1 + (g')^2 - gg''}{2g[1 + (g')^2]^{3/2}} \end{aligned}$$



It is also useful to consider the **canonical parameterization**, where $|c'|^2 = (c_1')^2 + (c_2')^2 = 1$. Then also $c_1'c_1'' + c_2'c_2'' = 0$, so we find

$$(4) \quad \begin{aligned} E &= 1 & F &= 0 & G &= c_1^2 \\ K &= \frac{c_2'(c_1'c_2'' - c_2'c_1'')}{c_1} = \frac{c_1'(c_2'c_2'') - c_1''(c_2')^2}{c_1} \\ &= \frac{c_1'(-c_1'c_1'') - c_1''[1 - (c_1')^2]}{c_1} \\ &= -\frac{c_1''}{c_1} \end{aligned}$$

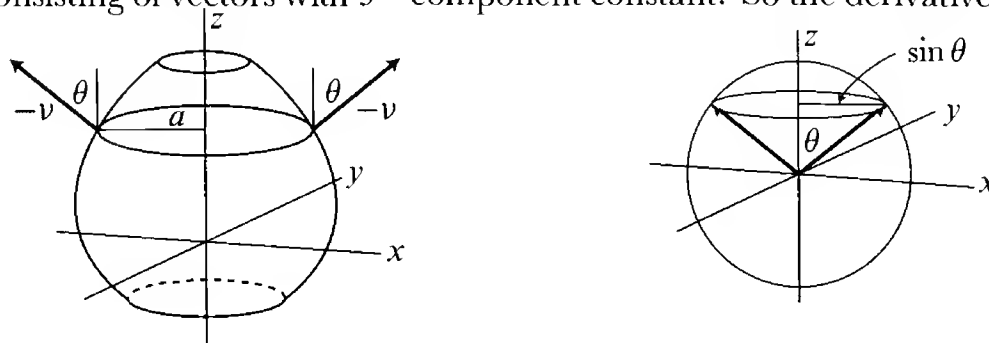
It is interesting to note that we can obtain all these results in a purely geometric way without any calculations. We first observe that the normals to the surface along the profile curve c lie in the (x, z) -plane. This means that



the derivative of ν along the profile curve also lies in the (x, z) -plane. Since it must also be tangent to the surface, it is a multiple of c' . Thus c' is one eigenvector for $-d\nu$. The corresponding eigenvalue is also easy to find. We notice first that our choice of N as $(f_1 \times f_2)/|f_1 \times f_2|$ makes ν inward pointing;

so along c , the vector v is just the normal \mathbf{n} of c . Therefore $-dv(c') = -\mathbf{n}' = \kappa \cdot c'$ (from the Serret-Frenet formulas for plane curves), and the eigenvalue is just the curvature of c . Comparing the formula for k_{meridian} in (2) with the formula on pg. II.8, we see that this is precisely what we have obtained in the calculations.

We next consider the outward pointing normal $-v$ along a parallel. This makes a constant angle θ with the z -axis, so in S^2 it traces out a circle, of radius $\sin \theta$, consisting of vectors with 3rd component constant. So the derivative of $-v$

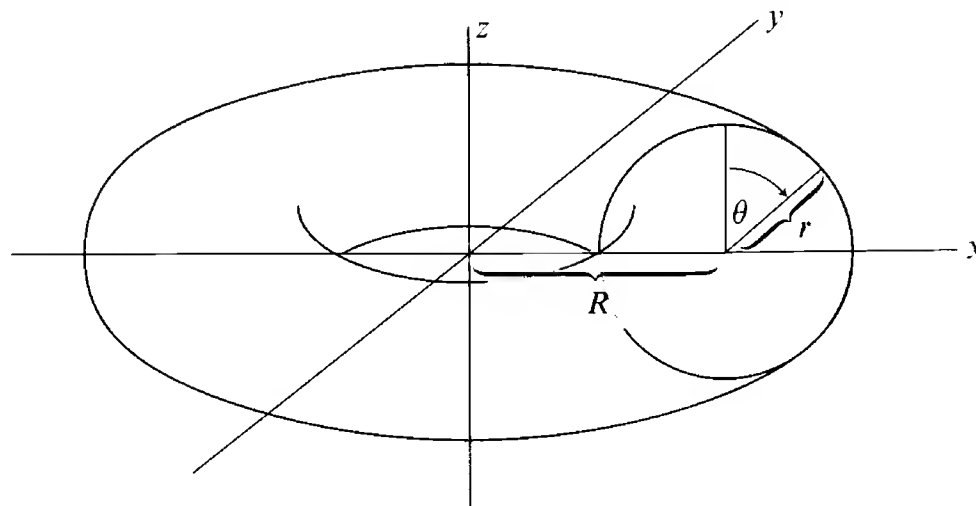


will be a vector with 3rd component 0, and perpendicular to the radius vector. It is therefore a multiple of the tangent vector of the parallel. If we parameterize our parallel so that we go once around in time 2π , then its tangent vector has length $a = \text{radius of the parallel}$. In the same time, the vector $-v$ goes once around a circle of radius $\sin \theta$, so its tangent vector has length $\sin \theta$. This shows that the corresponding eigenvalue is $\frac{1}{a} \sin \theta$, which is precisely what the formula for k_{parallel} in (2) gives.

Now let us take some particular cases. We begin with the most familiar example (after cylinders, cones, and spheres).

1. Torus

We rotate a circle of radius r around the z -axis so that its center traces out a circle of radius R . Measuring angles θ around the little circle clockwise from

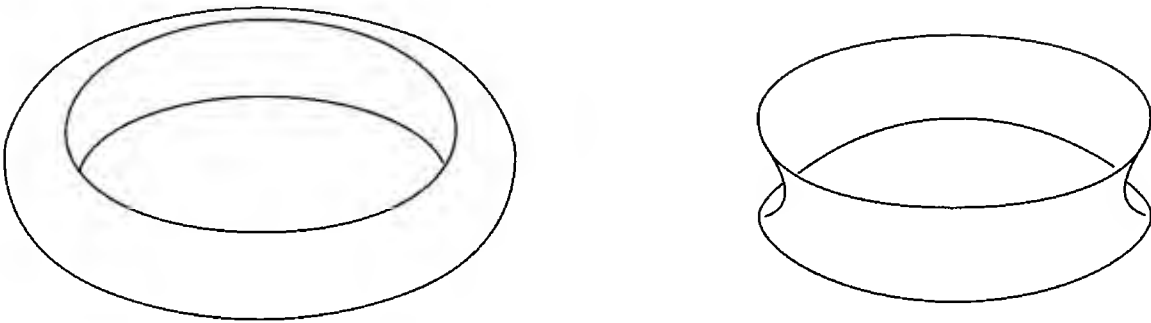


the z -axis, we see that the principal curvatures are:

$$\frac{1}{r} = \text{curvature of circle of radius } r$$

$$\frac{\sin \theta}{R + r \sin \theta}, \quad \text{since } a = R + r \sin \theta.$$

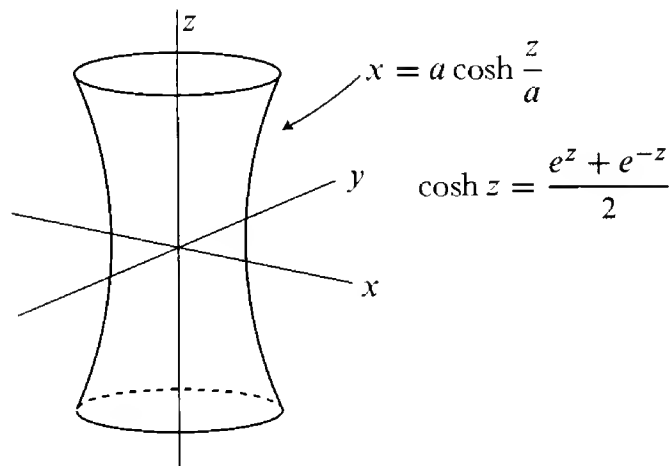
In particular, as suspected, $K > 0$ on the outer half of the torus, and $K < 0$ on



the inner half. Notice that the principal curvatures can never be equal, so there are no umbilics.

2. Catenoid

This surface, which we have already met in Volume I, Chapter 9, but not yet been formally introduced to, is obtained by revolving a **catenary**, with equation $x = a \cosh(z/a) = g(z)$, around the z -axis.



Since

$$\cosh' z = \frac{e^z - e^{-z}}{2} = \sinh z, \quad \cosh'' z = \frac{e^z + e^{-z}}{2} = \cosh z$$

$$1 + (\cosh' z)^2 = 1 + \frac{e^{2z} - 2 + e^{2z}}{4} = \frac{1}{2} + \frac{e^{2z}}{2} = (\cosh z)^2,$$

formulas (3) give us

$$k_{\text{meridian}} = \frac{-\frac{1}{a} \cosh \frac{z}{a}}{\left[\cosh^2 \frac{z}{a} \right]^{3/2}} = \frac{-1}{a \cosh^2 \frac{z}{a}}$$

$$k_{\text{parallel}} = \frac{1}{a \cosh \frac{z}{a} \left[\cosh^2 \frac{z}{a} \right]^{1/2}} = \frac{1}{a \cosh^2 \frac{z}{a}}$$

$$H = 0, \quad K = \frac{-1}{a^2 \cosh^4 \frac{z}{a}}.$$

Clearly $-1/a^2 \leq K < 0$, with $K = -1/a^2$ on the inner circle $z = 0$, and $K \rightarrow 0$ as $z \rightarrow \pm\infty$.

It is also useful to find the canonical parameterization for the catenoid; we take the case $a = 1$. For $c(u) = (\cosh u, u)$, we have

$$\text{length of } c \text{ from } 0 \text{ to } u = \int_0^u \sqrt{1 + (\cosh' v)^2} dv = \int_0^u \cosh v dv = \sinh u,$$

so we want to take the curve

$$\begin{aligned} \gamma(s) &= c(\sinh^{-1}(s)) = (\cosh(\sinh^{-1}(s)), \sinh^{-1}(s)) \\ &= (\sqrt{1+s^2}, \sinh^{-1}(s)) \quad [\text{see Problem I.9-20(d)}]. \end{aligned}$$

We then have, by formulas (4),

$$E = 1 \quad F = 0 \quad G = 1 + s^2$$

$$K = \frac{-1}{(1 + s^2)^2}.$$

3. Rotation Surfaces of Constant Curvature

We consider surfaces of revolution with canonical parameterization

$$f(s, t) = (g(s) \cos t, g(s) \sin t, h(s)), \quad (g')^2 + (h')^2 = 1,$$

and among these seek the ones with constant K . According to equations (4) we have

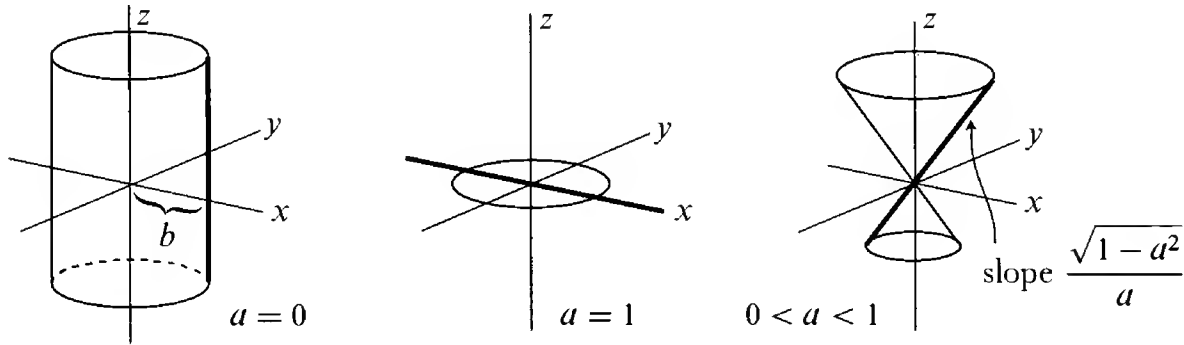
$$K = -\frac{g''}{g}.$$

Case 1. $K = 0$

Then $g(s) = as + b$. If $a \neq 0$, then we can assume that $b = 0$, since this merely amounts to renaming the parameter s . To have $(g')^2 + (h')^2 = 1$ we take

$$\left. \begin{aligned} g(s) &= as \\ h(s) &= \int_0^s \sqrt{1 - a^2} dt = \pm s\sqrt{1 - a^2} \end{aligned} \right\} \quad \text{or} \quad \begin{cases} g(s) = 0 \cdot s + b \\ h(s) = s \end{cases}$$

(changing h by a constant merely amounts to translating the profile curve along the z -axis); clearly we must have $|a| \leq 1$. For $a = 0$ we obtain a cylinder, for $|a| = 1$ a plane, and for $0 < |a| < 1$ a cone.



Case 2. $K > 0$

For simplicity, we take the case $K = 1$. We have to find g satisfying $g'' + g = 0$. The general solution, $g(s) = a_1 \cos s + a_2 \sin s$, can also be written

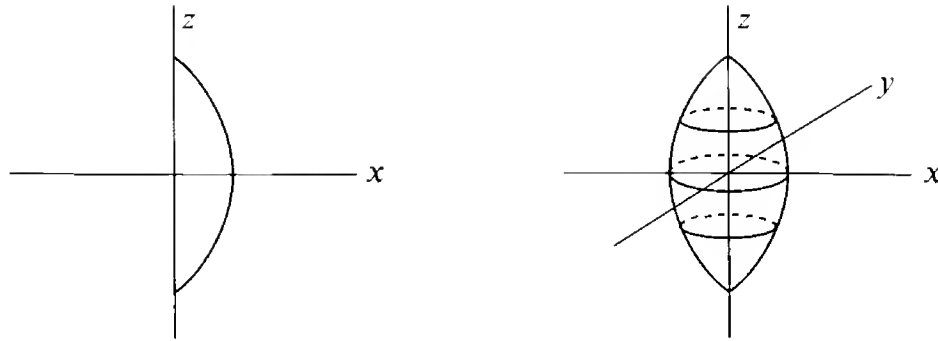
$$g(s) = a \cos(s + b).$$

We can always assume $b = 0$, and we take

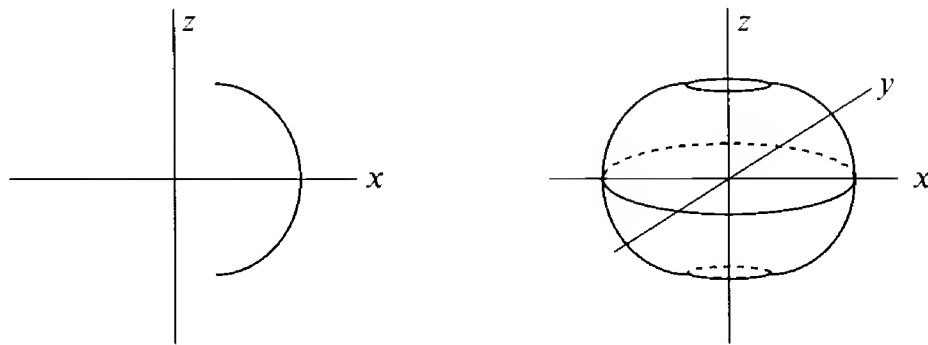
$$\begin{aligned} g(s) &= a \cos s \\ h(s) &= \pm \int_0^s \sqrt{1 - g'(t)^2} dt = \pm \int_0^s \sqrt{1 - a^2 \sin^2 t} dt. \end{aligned}$$

For $a = 1$ we obtain the sphere of radius 1.

For $a < 1$, the integrand in the expression for h is always real, and the only restriction on our formulas is that $g(s)$ must be ≥ 0 . We can take $0 \leq s < \pi/2$; the resulting profile curve can be expressed in terms of elliptic integrals.



For $a > 1$, we must restrict s to $0 \leq s \leq \arcsin 1/a$ for the integrand to be real. At the endpoint of the interval, h' is 0, so the profile curve has horizontal tangents; once again, elliptic integrals are involved.



Case 3. $K < 0$

We take the case $K = -1$. The general solution of $g'' - g = 0$ is

$$g(s) = ae^s + be^{-s}.$$

Suppose first that one of a, b is 0. We can assume that $b = 0$, since changing s to $-s$ interchanges a and b . We might as well assume $a > 0$, since changing a to $-a$ just changes the profile curve to its mirror image. Finally, we can assume $a = 1$, since changing s to $s + s_0$ multiplies a by e^{s_0} . So we take

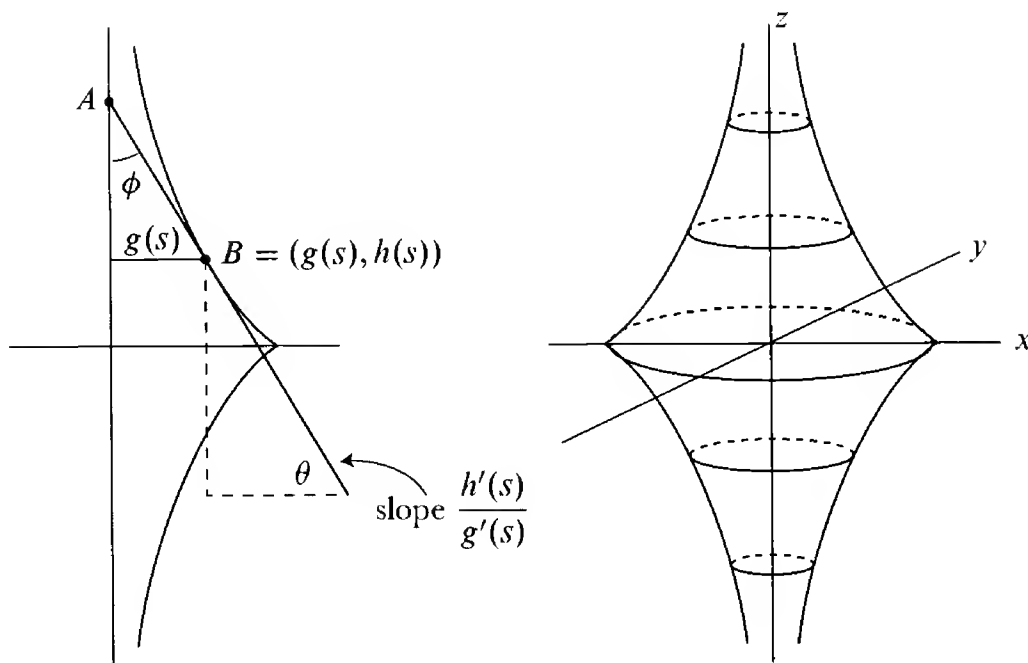
$$g(s) = e^s$$

$$h(s) = \pm \int_0^s \sqrt{1 - e^{2t}} dt;$$

we clearly need $e^{2s} \leq 1$, and therefore $g(s) \leq 1$. The resulting surface is called a **pseudosphere**. Its profile curve was known to mathematicians long before the advent of differential geometry. Since s is the parameterization by arclength,

the angle ϕ in the picture below satisfies

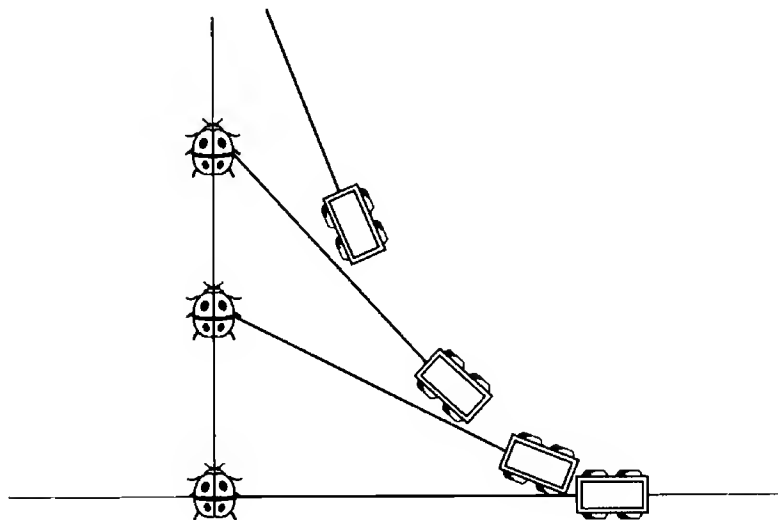
$$\sin \phi = \cos \theta = \frac{g'(s)}{\sqrt{[g'(s)]^2 + [h'(s)]^2}} = g'(s) = e^s.$$



So AB has constant length

$$\overline{AB} = \frac{g(s)}{\sin \phi} = \frac{e^s}{e^s} = 1.$$

If one started at $(0,0)$ and walked along the y -axis pulling a wagon that started at $(1,0)$ and had a handle of length 1, then the wagon would follow this curve, which is therefore called a **tractrix** (Latin: *trahere*, *tractum* to draw). The upper



half of the tractrix is the graph of the function

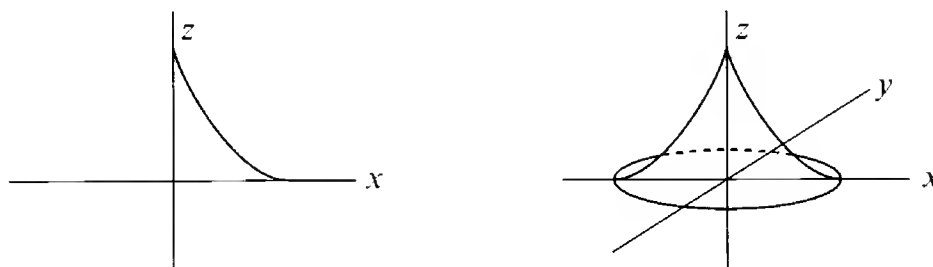
$$\begin{aligned} f(x) &= \int_0^{\log x} \sqrt{1 - e^{2t}} \, dt \\ &= \int_1^x \sqrt{1 - u^2} \cdot \frac{1}{u} \, du \\ &= \sqrt{1 - x^2} - \cosh^{-1} \frac{1}{x}. \end{aligned}$$

Now suppose that $a, b \neq 0$. Since changing s to $s + s_0$ multiplies a and b by different constants, we can assume that either $a = -b$ or $a = b$.

In the case $a = -b$ we can assume $a > 0$ (by changing s to $-s$ and thereby interchanging a and b). So we take

$$\begin{aligned} g(s) &= a(e^s - e^{-s}) = 2a \sinh s \\ h(s) &= \pm \int_0^s \sqrt{1 - 4a^2 \cosh^2 t} \, dt; \end{aligned}$$

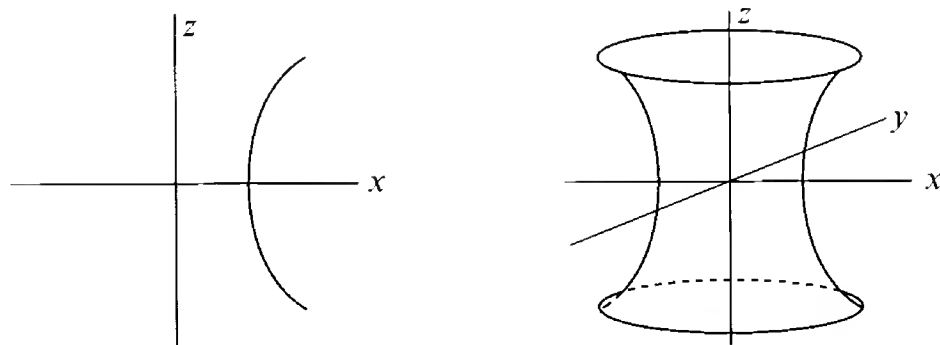
we need $0 < 2a < 1$ and $1 \leq \cosh s \leq 1/2a$, so that $0 \leq s \leq \cosh^{-1} 1/2a$ and $0 \leq g(s) \leq \sqrt{1 - 4a^2}$. These functions can also be expressed in terms of elliptic integrals.



In case $a = b$, we can assume both are positive, since changing the sign of both changes the profile curve to its mirror image. So we take

$$\begin{aligned} g(s) &= 2a \cosh s \\ h(s) &= \pm \int_0^s \sqrt{1 - 4a^2 \sinh^2 t} \, dt; \end{aligned}$$

we need $|\sinh s| \leq 1/2a$ and thus $2a \leq g(s) \leq \sqrt{1 + 4a^2}$. Elliptic integrals are again required.

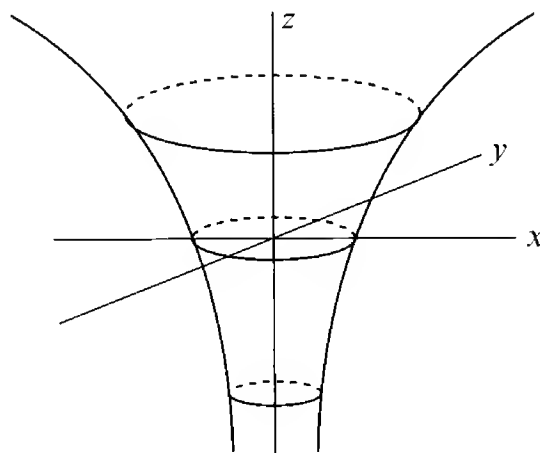


These results about surfaces of revolution may be compared with the remarks made by Riemann in section II.5 of his Inaugural Lecture (pg. II.159).

4. A Classical Counterexample

Consider the surface of revolution

$$f(s, t) = (s \sin t, s \cos t, \log s).$$



Formulas (1) and (2) give

$$\begin{aligned} E &= 1 + \frac{1}{s^2} & F &= 0 & G &= 1 \\ k_{\text{meridian}} &= \frac{-1/s^2}{\left[1 + \frac{1}{s^2}\right]^{3/2}} & k_{\text{parallel}} &= \frac{1/s}{s \left[1 + \frac{1}{s^2}\right]^{1/2}} \\ K &= \frac{-1}{(1 + s^2)^2}. \end{aligned}$$

On the other hand, interchanging the role of s and t in the parameterization of the helicoid (page 150), we see that $g(s, t) = (s \cos t, s \sin t, t)$ has

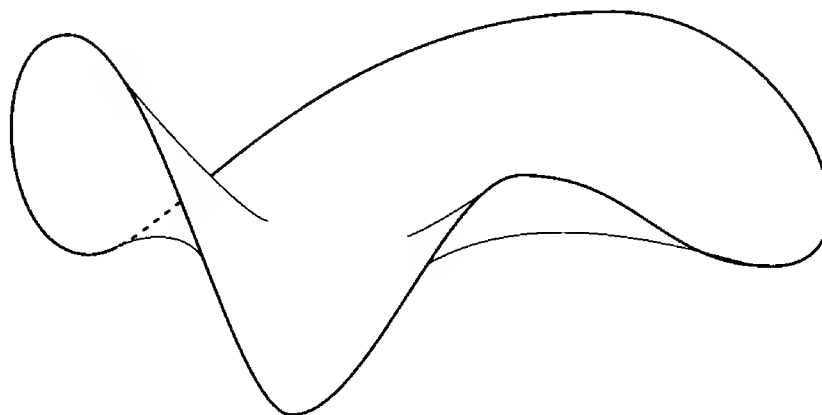
$$E = 1 \quad F = 0 \quad G = 1 + s^2$$

$$K = \frac{-1}{(1 + s^2)^2}.$$

Consequently, the map $f(s, t) \mapsto g(s, t)$ preserves K , but is not an isometry; in fact, there is clearly no local isometry between the two surfaces, since the s -parameters would have to correspond to preserve K , and then E would not be preserved.

V. MINIMAL SURFACES

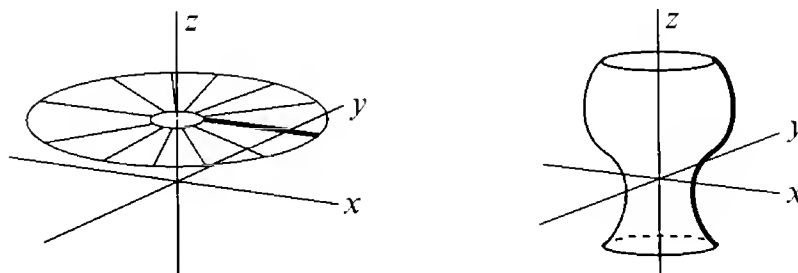
A whole branch of mathematics is devoted to the study of surfaces with mean curvature $H = 0$. As we shall show in Chapter 9, this condition is precisely the one which must be satisfied by a surface which is a critical point for the area function among all surfaces with the same boundary curve c . In particular, if a



surface has minimum area among those with c as boundary, then it must satisfy the condition $H = 0$. For this reason, surfaces with mean curvature $H = 0$ are called **minimal**.

We have already met two minimal surfaces in our survey, the helicoid and the catenoid. They, in fact, were the first two non-planar minimal surfaces to be discovered (by Meusnier), and we are led to them directly if we seek minimal surfaces among other known classes of surfaces.

Let us consider first the surfaces of revolution. If our profile curve is a straight line perpendicular to the z -axis, we obtain a plane. Otherwise, some portion of the curve can be represented by $c(s) = (g(s), s)$. Using formulas (3) on



page 158, we find that $H = 0$ when

$$1 + (h')^2 - hh'' = 0.$$

This is precisely the equation we obtained on pg. I.321, when we were actually finding the minimum area of a surface of revolution. We found there that the solutions are

$$g(z) = a \cosh \frac{z+b}{a}.$$

This result applies only to a portion of the surface, but a simple least upper bound argument shows that if a connected profile curve has this form somewhere, then it must have it everywhere (and in particular cannot also contain part of a line perpendicular to the z -axis). We have thus shown that:

Any connected minimal surface of revolution is part of a plane or a catenoid.

We consider next the ruled surface

$$\begin{aligned} f(s, t) &= \sigma(s) + t\delta(s) \\ |\delta| &= |\delta'| = 1, \quad \langle \sigma', \delta' \rangle = 0 \\ [\delta, \sigma' &\text{ linearly independent}]. \end{aligned}$$

Then we have

$$\begin{aligned} f_1 &= \sigma' + t\delta' & f_2 &= \delta \\ f_{11} &= \sigma'' + t\delta'' & f_{22} &= 0 \\ f_{12} &= \delta' \\ F &= \langle \delta, \sigma' \rangle & G &= 1; \quad \text{let } W = \sqrt{EG - F^2} \\ l &= \frac{1}{W} \det \begin{pmatrix} \sigma'' + t\delta'' \\ \sigma' + t\delta' \\ \delta \end{pmatrix} & m &= \frac{1}{W} \det \begin{pmatrix} \delta' \\ \sigma' + t\delta' \\ \delta \end{pmatrix} & n &= 0. \end{aligned}$$

So equations (B) show that $H = 0$ when

$$0 = -2\langle \delta, \sigma' \rangle \det \begin{pmatrix} \delta' \\ \sigma' + t\delta' \\ \delta \end{pmatrix} + \det \begin{pmatrix} \sigma'' + t\delta'' \\ \sigma' + t\delta' \\ \delta \end{pmatrix}.$$

Using multilinearity of \det as a function of its rows, and noting that the coefficient of each power of t must vanish, we obtain

$$(1) \quad \langle \delta, \sigma' \rangle \det \begin{pmatrix} \delta' \\ \sigma' \\ \delta \end{pmatrix} = 0 \quad (2) \quad \det \begin{pmatrix} \sigma'' \\ \delta' \\ \delta \end{pmatrix} + \det \begin{pmatrix} \delta'' \\ \sigma' \\ \delta \end{pmatrix} = 0 \quad (3) \quad \det \begin{pmatrix} \delta'' \\ \delta' \\ \delta \end{pmatrix} = 0.$$

Equation (3) shows that δ'' is a linear combination of δ and δ' . But we also have

$$\langle \delta', \delta' \rangle = 1 \implies \langle \delta', \delta'' \rangle = 0$$

$$\langle \delta, \delta \rangle = 1 \implies \langle \delta, \delta' \rangle = 0 \implies \langle \delta', \delta' \rangle + \langle \delta, \delta'' \rangle = 0 \implies \langle \delta, \delta'' \rangle = -1,$$

which shows that

$$\delta'' = -\delta.$$

This means that $-\delta$ is the normal \mathbf{n} of the curve δ , and that the curvature of δ is $\kappa = 1$. Also, $\mathbf{b} = \mathbf{t} \times \mathbf{n} = \delta' \times -\delta$, so

$$\mathbf{b}' = -(\delta' \times \delta)' = -(\delta' \times \delta') - (\delta'' \times \delta) = 0.$$

Thus $\tau = 0$, and δ is a plane curve. Since it lies in S^2 , and has curvature 1, it must be a circle of radius 1. We can assume therefore that

$$\delta(s) = (\cos s, \sin s, 0).$$

Now in equation (2), the second determinant is already 0, so we find that σ'' is a linear combination of δ, δ' , which means that σ'' lies in the (x, y) -plane. So σ must be of the form

$$\sigma(s) = (\alpha(s), \beta(s), bs + a) \implies \sigma'(s) = (\alpha'(s), \beta'(s), b).$$

We might as well assume that $a = 0$, since this just amounts to a translation along the z -axis.

Now consider equation (1), which says that a certain product is 0. If the second factor is 0 for some s_0 , then $\sigma'(s_0)$ must be a linear combination of $\delta(s_0), \delta'(s_0)$, so $b = 0$. In this case, all rulings $\sigma(s) + t\delta(s)$ lie in a plane, and our surface is just the plane. If $b \neq 0$, then for all s we must have

$$0 = \langle \delta(s), \sigma'(s) \rangle = \alpha'(s) \cos s + \beta'(s) \sin s$$

$$0 = \langle \delta'(s), \sigma'(s) \rangle = -\alpha'(s) \sin s + \beta'(s) \cos s.$$

So $\alpha' = \beta' = 0$, i.e., α and β are constants, and our surface is given by

$$f(s, t) = (\alpha + t \cos s, \beta + t \sin s, bs).$$

A translation in the (x, y) -plane changes α and β to 0, and we obtain the helicoid.

Our analysis has left a few points in doubt, because the standard parameterization with which we began is possible only when the directions of the rulings of our ruled surface are always changing. When the directions of the rulings are *never* changing we obtain a generalized cylinder, which is minimal only if it is a plane. As before, a least upper bound argument shows that if $\delta' \neq 0$ on some interval, so that we do have a helicoid on this interval, then $\delta' \neq 0$ everywhere. We have thus shown that:

Any connected minimal ruled surface is part of a plane or a helicoid.

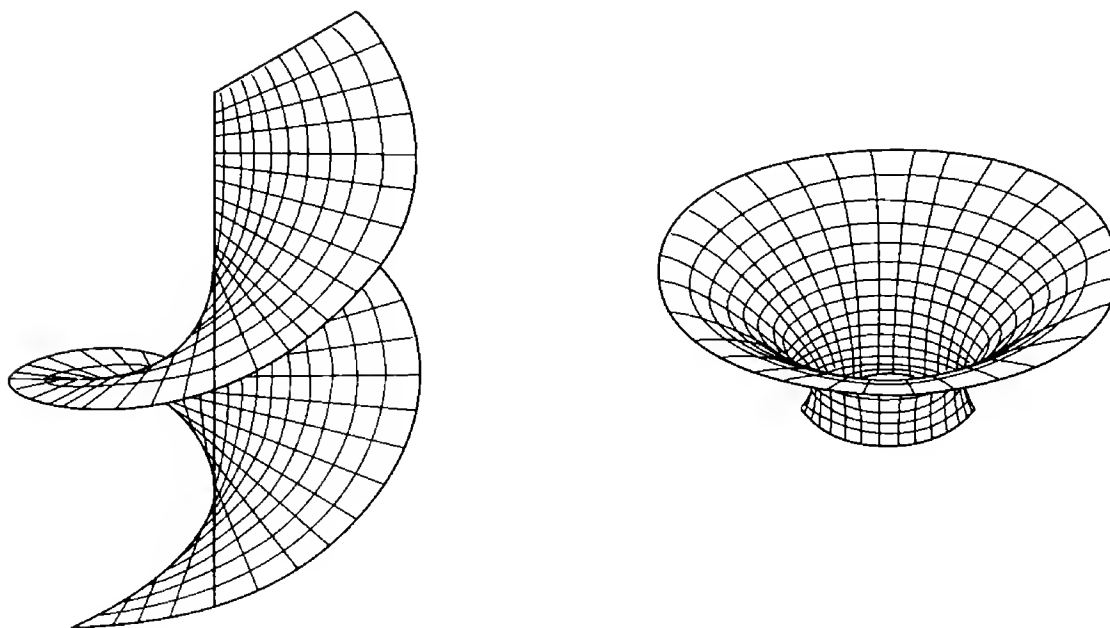
It is, of course, not particularly surprising that each of these families of surfaces contains only one non-planar minimal surface. But it is rather surprising that these two surfaces, *the catenoid and the helicoid, are locally isometric*. To prove this, we merely recall (page 167) that

the helicoid $f(s, t) = (s \cos t, s \sin t, t)$ has $E = 1 \quad F = 0 \quad G = 1 + s^2$,
while (page 161)

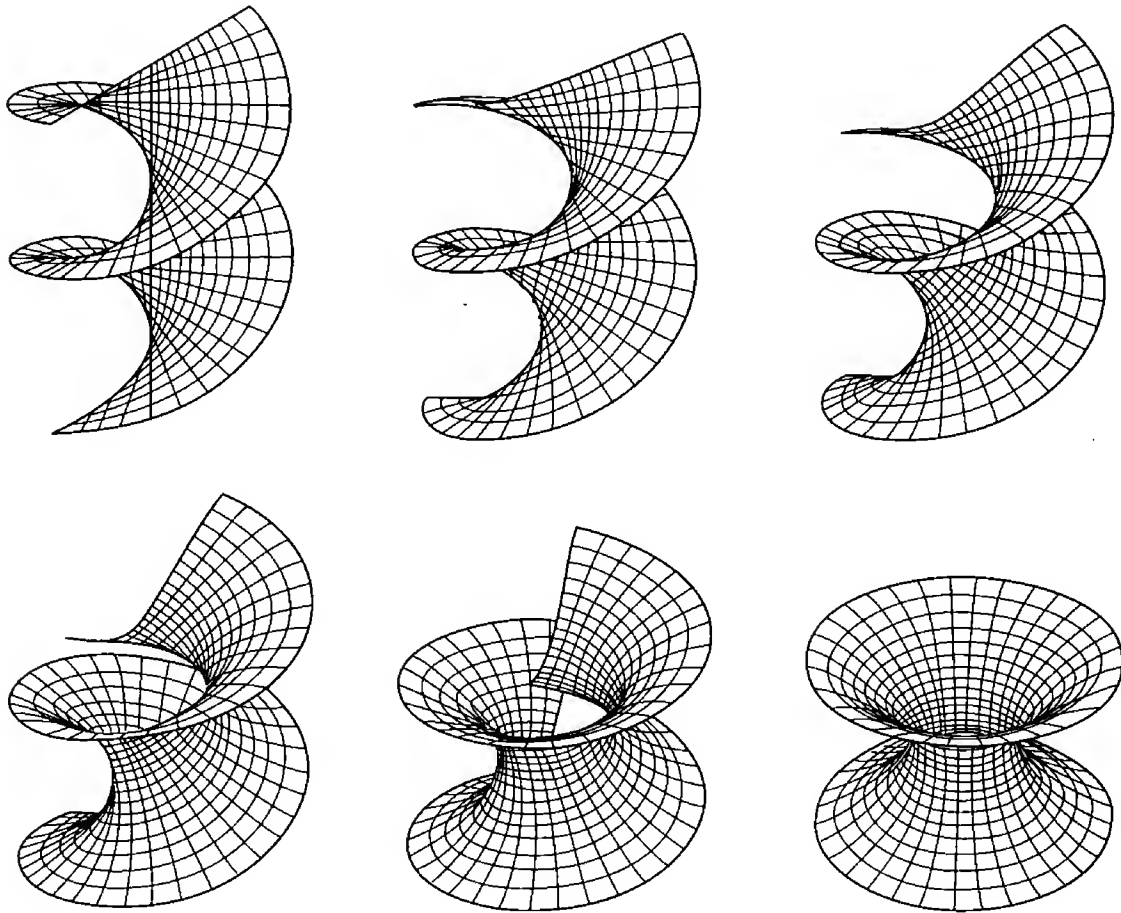
the catenoid $g(s, t) = (\sqrt{1 + s^2} \cos t, \sqrt{1 + s^2} \sin t, \sinh^{-1}(s))$ also has
 $E = 1 \quad F = 0 \quad G = 1 + s^2$.

The isometry, taking $f(s, t)$ to $g(s, t)$, carries

rulings of the helicoid to *meridians* of the catenoid (t constant)
helices of the helicoid to *parallels* of the catenoid ($s \neq 0$ constant)
z-axis of the helicoid to *center circle* of the catenoid ($s = 0$).



Actually, we can do a lot better than this: it is possible to deform one of these surfaces into the other by means of a 1-parameter family of isometric surfaces:



Although we could write down an explicit formula for this 1-parameter family, it will come out very naturally in Chapter 9.

The example of the helicoid and catenoid also shows that two immersions, $f, \tilde{f}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ with the same g_{ij} (and therefore also the same K) as well as the same H need not differ by a Euclidean motion of \mathbb{R}^3 —one needs to know the l_{ij} themselves, not just $\text{trace}(l_{ij})$ and $\det(l_{ij})$ [or equivalently the eigenvalues of (l_{ij})], in order to determine the surface.

The first minimal surface to be discovered after the catenoid and helicoid was

1. *Scherk's Minimal Surface*

This is the surface M defined by

$$e^z \cos x = \cos y.$$

If we define $W: \mathbb{R}^3 \rightarrow \mathbb{R}$ by

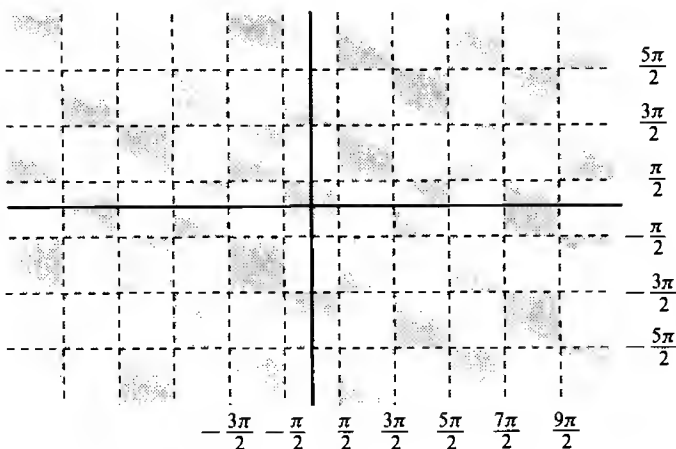
$$W(x, y, z) = e^z \cos x - \cos y,$$

then the three vectors

$$W_1(x, y, z) = -e^z \sin x \quad W_2(x, y, z) = \sin y \quad W_3(x, y, z) = e^z \cos x$$

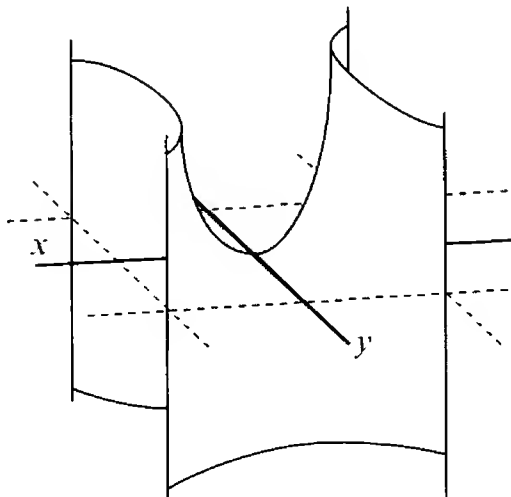
are never all 0, so $W^{-1}(0)$ is a surface, which is orientable since it has a well-defined normal, as on page 138. The lines $x = \pi/2 + m\pi$ and $y = \pi/2 + m\pi$, for $m \in \mathbb{Z}$, divide \mathbb{R}^2 into squares, and those where $\cos x \cos y > 0$ form a checkerboard pattern. Since $e^z > 0$, there are clearly no points of M over the

1. P-Q4, N-KB3
2. N-Q2, P-K4
3. P×P, N-N5
4. P-KR3, N-K6
5. Resign

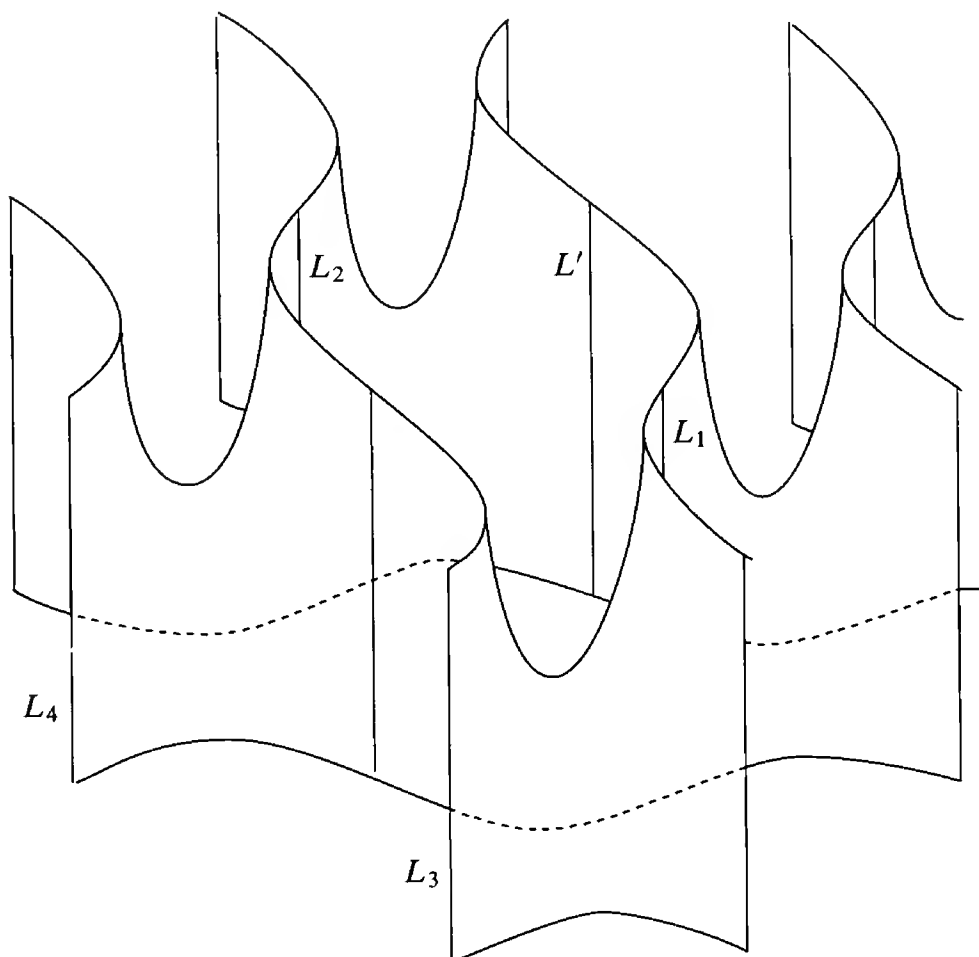


“white” squares. Over the vertices of the squares, where $\cos x = \cos y = 0$, we have perpendicular lines, since z can have any value. Over the “black” squares, we can solve explicitly for z ,

$$z = \log \left(\frac{\cos y}{\cos x} \right).$$



The surface is made up of infinitely many of these structural units.



The small portion of the surface pictured above is simply homeomorphic to a cylinder (with a very floppy side), and there is an obvious closed curve c connecting the four saddle points which represents a non-trivial element of the fundamental group. Moreover, this curve c does not disconnect the (complete) surface. To see this, note that just as the vertical line L_1 can be connected to the line L_2 by a curve lying in the two adjacent units which share the common edge L' , so lines L_3 and L_4 can be connected by a curve lying in two adjacent units (not drawn in the picture). But L_3 and L_4 can be connected to points which lie on (apparently) opposite sides of the curve c .

Similarly, we easily see that the surface actually has “infinite genus” (infinitely many closed curves may be removed from it without disconnecting it). It also has only one end. But it is known that two orientable surfaces with the same genus and the same space of ends are homeomorphic, so our surface must be homeomorphic to surface (A) in Problem I.1-20.

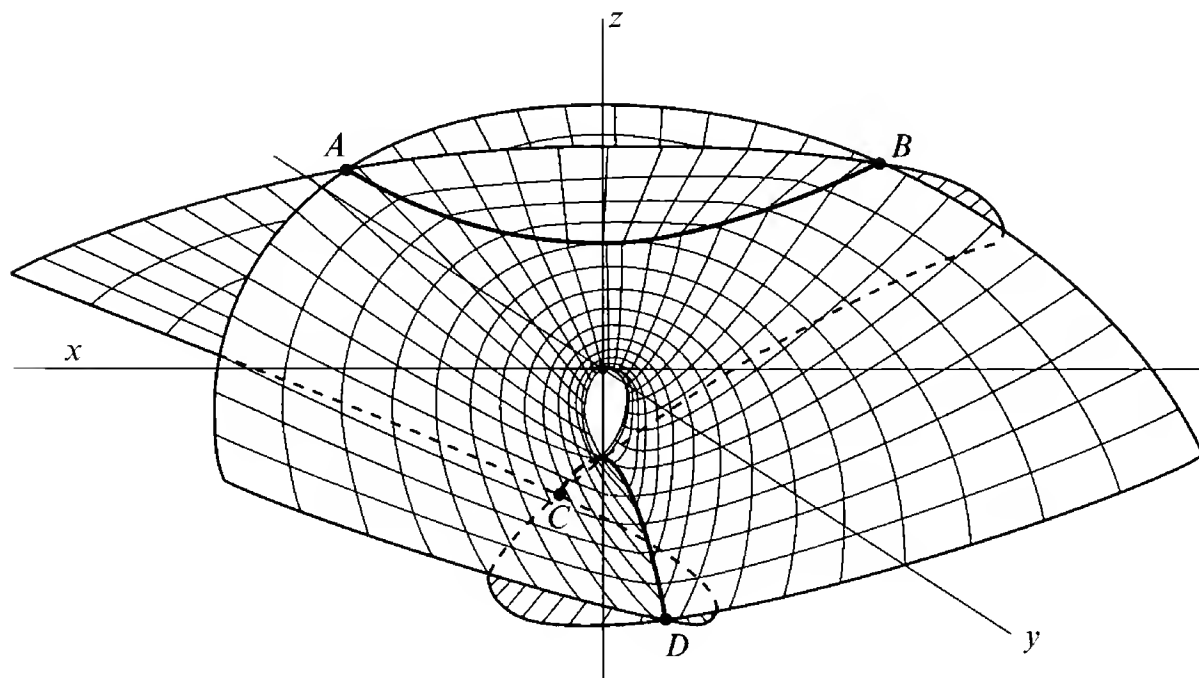
2. Enneper's Minimal Surface

This surface is parameterized by

$$f(u, v) = \left(\frac{u}{2} - \frac{u^3}{6} + \frac{uv^2}{2}, -\frac{v}{2} + \frac{v^3}{6} - \frac{u^2v}{2}, \frac{u^2}{2} - \frac{v^2}{2} \right).$$

A computation from equations (A) and (B) shows that $H = 0$. Of course, at present it is pretty hard to see how any one ever thought of this example, but in Chapter 9 we will see that in a certain sense it is the simplest minimal surface.

In the figure below, showing the image of f on $[-2.5, 2.5] \times [-2.5, 2.5]$, the top portion is seen almost completely from the side, while the bottom portion

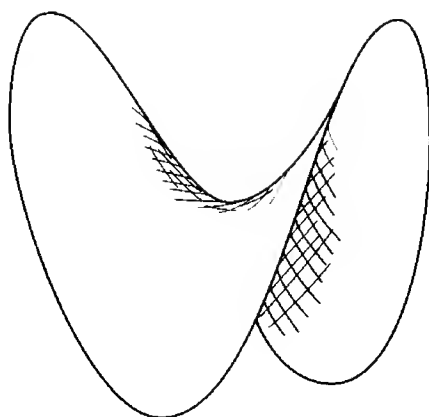


is seen almost head on. The surface is taken into itself by the map

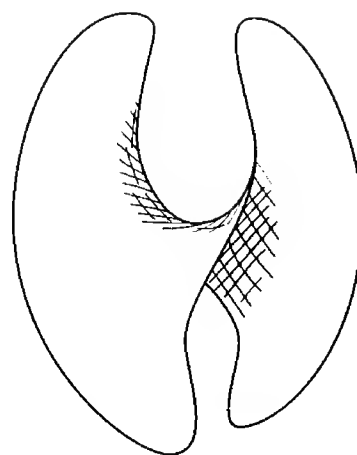
$$(x, y, z) \mapsto (y, x, -z),$$

with the line AB corresponding to the line CD .

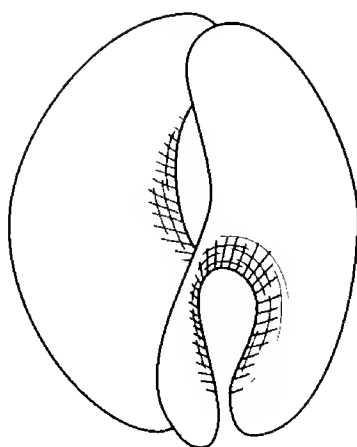
Since the surface is the graph of a function, it is merely an immersed plane, and its structure can perhaps be better understood from the series of pictures on the next page, which show a saddle surface being deformed into a surface of the same type as Enneper's surface.



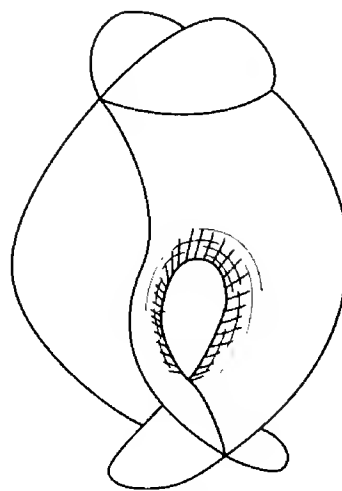
(a)



(b)



(c)



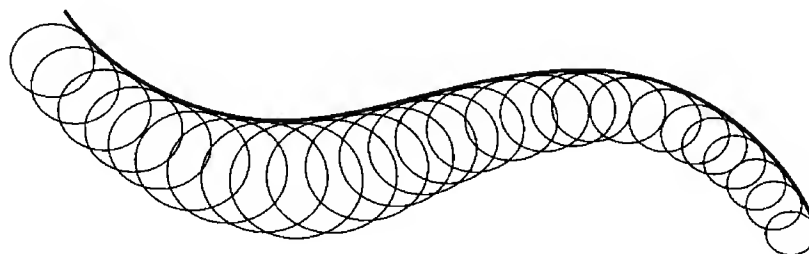
(d)

ADDENDUM

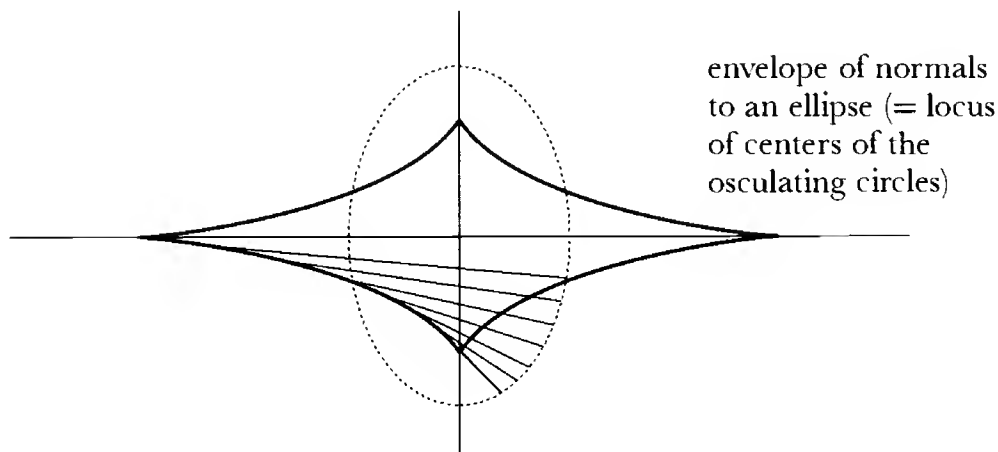
ENVELOPES OF 1-PARAMETER FAMILIES OF PLANES

In classical differential geometry, a central role was played by the notion of the envelope of a family of curves or surfaces. A careful treatment of this topic involves many delicate points, which to be sure were rather indelicately handled by classical geometers. However, the study of envelopes played such an important role in the evolution of the concept of a connection that a sketch of its essential features seems in order; the ideas developed here will also be used on a couple of later occasions.

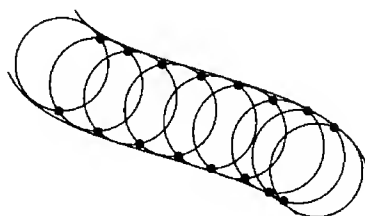
Consider first a 1-parameter family $\tilde{\alpha}$ of curves in the plane, given by $\tilde{\alpha}(u) = t \mapsto \alpha(u, t)$ for some C^∞ function $\alpha: [0, 1] \times [0, 1] \rightarrow \mathbb{R}^2$. An **envelope** of this family is defined to be a curve c which is *not* a member of this family but which is tangent to some member of the family at every point. Unfortunately, it often



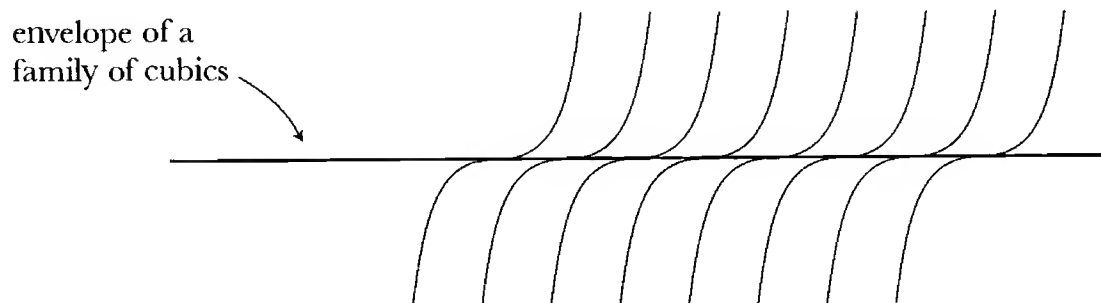
turns out that the envelope of a perfectly nice family of curves has a cusp or something worse; but for the time being we won't worry too much about this.



The classical way of finding the envelope of α was very geometric. For each u , we let $c(u)$ be the limit, as $\varepsilon \rightarrow 0$, of the intersection of $\bar{\alpha}(u)$ and $\bar{\alpha}(u + \varepsilon)$: the

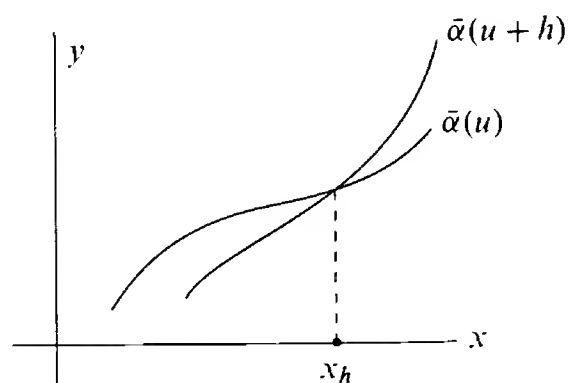


envelope consists of the “intersections of members of the family with another member infinitely close to it”. The picture below shows that this idea can run into some serious difficulties. Nevertheless, it often works out rather well in



particular cases, and even in the general case it leads us to the proper analytic condition, when we argue as follows.

Let us consider first the case where our curves $\bar{\alpha}(u)$ are all expressed as the graphs of functions; thus there is a function $(u, x) \mapsto f(u, x)$ such that $\alpha(u, t) = (t, f(u, t))$. Suppose that the curve $\bar{\alpha}(u)$ intersects the curve $\bar{\alpha}(u+h)$ at the point



$$(x_h, f(u, x_h)) = (x_h, f(u + h, x_h)).$$

Then we have

$$0 = \frac{f(u + h, x_h) - f(u, x_h)}{h}.$$

Assuming that x_h approaches a number $x(u)$ as $h \rightarrow 0$, we find that $x(u)$ must be a point for which

$$(*) \quad D_1 f(u, x(u)) = 0.$$

If we find the points $x(u)$ for all u , then the envelope should be the curve consisting of all points $(x(u), f(u, x(u)))$.

If we are given a general family $\bar{\alpha}$, not necessarily expressed as graphs of functions, then we can introduce the function f in two steps. We first determine $t(u, x)$ so that

$$(1) \quad \alpha^1(u, t(u, x)) = x,$$

and then define

$$(2) \quad f(u, x) = \alpha^2(u, t(u, x)).$$

Then equation $(*)$ becomes

$$(3) \quad 0 = D_1 \alpha^2(u, t(u, x)) + D_2 \alpha^2(u, t(u, x)) \cdot D_1 t(u, x),$$

while equation (1) gives

$$\begin{aligned} D_1 \alpha^1(u, t(u, x)) + D_2 \alpha^1(u, t(u, x)) \cdot D_1 t(u, x) &= 0, \\ D_1 t(u, x) &= -\frac{D_1 \alpha^1}{D_2 \alpha^1}(u, t(u, x)). \end{aligned}$$

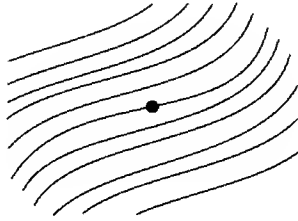
Substituting this into (3), we obtain

$$[D_1 \alpha^2 \cdot D_2 \alpha^1 - D_1 \alpha^1 \cdot D_2 \alpha^2](u, t(u, x)) = 0.$$

Thus we find that the envelope should consist of points $\alpha(u, t)$ where (u, t) satisfies

$$(**) \quad \det(D_i \alpha^j(u, t)) = 0.$$

Now even without resorting to the motivating geometric construction, it is clear that if there is an envelope of the family $\bar{\alpha}$, then it must be a subset of the points $\alpha(u, t)$ for which (u, t) satisfies (**). For, if the determinant in (**) is non-zero, then α is an immersion at (u, t) , and the curves $\bar{\alpha}(u)$ form a foliation of a neighborhood of $\alpha(u, t)$; consequently, the only curve through $\alpha(u, t)$ which



is tangent to some curve of the family at each point is $\bar{\alpha}(u)$ itself, which means that $\alpha(u, t)$ cannot be a point of an envelope.

Similar considerations will apply to 1-parameter families of surfaces in space. In particular, the geometric construction will be found quite useful when we consider 1-parameter families of planes. A plane P can be described as

$$\{x \in \mathbb{R}^3 : a_1 x_1 + a_2 x_2 + a_3 x_3 = \langle a, x \rangle = d\}$$

for some number d , and some $(a_1, a_2, a_3) \neq 0$, which we might as well assume is a unit vector. (Choosing the point $(x_1, x_2, x_3) \in P$ closest to 0, and noting that it must be a multiple of a , we see that d is just the distance from 0 to P , provided that a is picked so that it points in the direction of points further from 0). So a 1-parameter family of planes amounts to two functions $a: \mathbb{R} \rightarrow S^2$ and $d: \mathbb{R} \rightarrow \mathbb{R}$. Obviously if $a' = 0$, so that a is constant, then we obtain a family of parallel planes, and there is no envelope. Let us assume that, in fact, a' is *never* 0. Then any two nearby planes, corresponding to $u < v$, must intersect in a straight line, and all x on this line satisfy

$$\sum_i a_i(u)x_i - d(u) = \sum_i a_i(v)x_i - d(v) = 0.$$

Applying Rolle's Theorem to

$$s \mapsto \sum_i a_i(u + s[v - u])x_i - d(u + s[v - u]) \quad \text{on } [0, 1],$$

we find that

$$(1) \quad \sum_i a_i'(\xi)x_i - d'(\xi) = 0$$

for some ξ (depending on x) between u and v .

It follows that as $v \rightarrow u$, this line approaches the line satisfying

$$(*) \quad \begin{cases} \langle a(u), x \rangle = d(u) \\ \langle a'(u), x \rangle = d'(u) \end{cases}$$

(note that $a \in S^2$ implies that a' is perpendicular to a , so the planes given by these two equations are not parallel).

The line determined by $(*)$ is called the **characteristic line** of the plane determined by u ; its direction is $a(u) \times a'(u)$. If $(a \times a')' = 0$, so that all characteristic lines are parallel, then the envelope will be the generalized cylinder formed by all these characteristic lines, provided this actually exists (it could happen, for example, that all characteristic lines are the same, in which case no envelope would exist). Let us assume that, in fact, $(a \times a')'$ is *never* 0. Then any three nearby planes, corresponding to $u < v < w$, must have linearly independent $a(u), a(v), a(w)$, so the planes corresponding to u, v, w must intersect at a point (x_1, x_2, x_3) . This point must satisfy (I) and an analogous condition for v, w :

$$\sum_i a_i'(\bar{\xi})x_i - d'(\bar{\xi}) = 0$$

for some $\bar{\xi}$ between v and w .

Arguing as before, we see that

$$\sum_i a_i''(\eta)x_i - d''(\eta) = 0$$

for some η between ξ and $\bar{\xi}$.

It follows that as $v, w \rightarrow u$, this point approaches the point $c(u)$ satisfying

$$(**) \quad \begin{cases} \langle a(u), c(u) \rangle = d(u) \\ \langle a'(u), c(u) \rangle = d'(u) \\ \langle a''(u), c(u) \rangle = d''(u). \end{cases}$$

The point $c(u)$ determined by $(**)$ is called the **characteristic point** of the plane determined by u , and lies on its characteristic line. It is possible that $c' = 0$, so that c is a point. In this case, the envelope is the generalized cone formed by all the characteristic lines with c as vertex. Let us assume that, in fact, c' exists and is *never* 0. Differentiating the first two equations of $(**)$ gives

$$\begin{aligned} \langle a'(u), c(u) \rangle + \langle a(u), c'(u) \rangle &= d'(u), & \text{hence} \quad (2) \quad \langle a(u), c'(u) \rangle &= 0 \\ \langle a''(u), c(u) \rangle + \langle a'(u), c'(u) \rangle &= d''(u), & \text{hence} \quad (3) \quad \langle a'(u), c'(u) \rangle &= 0. \end{aligned}$$

Differentiating (2) then gives

$$\langle a'(u), c'(u) \rangle + \langle a(u), c''(u) \rangle = 0,$$

which together with (3) gives

$$(4) \quad \langle a(u), c''(u) \rangle = 0.$$

We thus have:

- $c(u)$ is the characteristic point of the plane determined by u
- $c'(u)$ has the same direction as the characteristic line (*)
- [by (2) and (3)]
- $c''(u)$ is in a plane parallel to the plane determined by u
- [by (4)].

So the plane determined by u is the osculating plane of c , and the tangent developable of c is the envelope of the family. Each plane of the family is tangent to this developable along the points where it intersects it, namely along its characteristic line. (For all this to work, of course, we need c'' to be non-zero.)

To sum things up, a 1-parameter family of planes “in general” has an envelope, which is either a generalized cylinder, a generalized cone, or the tangent developable of a curve. This well-known fact from classical differential geometry was precisely what gave Levi-Civita the clue for defining parallel translation in a Riemannian-manifold. He first observed that since generalized cylinders, generalized cones, and tangent developables are locally isometric to the plane, it makes sense to talk about parallel vector fields in these surfaces—they are just the images, under the local isometry, of parallel vector fields in the plane.

Now suppose that we are given a curve c on an arbitrary surface M . Consider the 1-parameter family of planes formed by the tangent planes $M_{c(u)}$. This family “generally speaking” has an envelope N , which is a generalized cylinder or cone, or the tangent developable to some (other) curve; and the tangent space of N is the same as that of M all along the curve c . So we can define a vector field V_u to be parallel along c in M if it is parallel along c in N . Once Levi-Civita had this definition of parallel vector fields along a curve c in M it was not hard to derive the usual equation for it, in terms of the Christoffel symbols. This equation shows that parallelism does not depend on the imbedding, and can be used to define parallel vector fields along a curve in any Riemannian manifold, of any dimension.

PROBLEMS

1. If $a_1 f_1 + a_2 f_2$ is a principal vector, then

$$\begin{pmatrix} G & -F \\ -F & E \end{pmatrix} \begin{pmatrix} l & m \\ m & n \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

for some λ . Write the two equations which this gives, and divide by a_1 and a_2 , respectively, to obtain two expressions for λ . From this derive equation (D) (and then check that it holds also for a_1 or $a_2 = 0$).

2. Let $f: V \rightarrow \mathbb{R}$ be a linear function on a (possibly infinite-dimensional) vector space V .

(a) If $v_1, v_2 \in V$, then $f(v_1)v_2 - f(v_2)v_1 \in \ker f$.

(b) We can write $V = \ker f \oplus W$, where W is 1-dimensional.

(c) If $g: V \rightarrow \mathbb{R}$ is a linear function with $\ker f \subset \ker g$, then $g = \lambda f$ for some $\lambda \in \mathbb{R}$.

(d) If $g, f_1, \dots, f_k: V \rightarrow \mathbb{R}$ and $\bigcap_i \ker f_i \subset \ker g$, then $g = \sum_i \lambda_i f_i$ for some $\lambda_i \in \mathbb{R}$.

3. Let the Jacobian of $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$ have rank k on $f^{-1}(0)$, so that $M = f^{-1}(0)$ is a submanifold of \mathbb{R}^n of dimension $n - k$. Let $g: M \rightarrow \mathbb{R}$ be differentiable, and suppose that g has a maximum at $p \in M$.

(a) $M_p = \bigcap_{i=1}^k \ker df^i$, where $df^i: \mathbb{R}^n_p \rightarrow \mathbb{R}$.

(b) If $X_p \in M_p$ then $dg(X_p) = 0$. *Hint:* $X_p = c'(0)$ for some curve c in M .

(c) Use Problem 2 to conclude that there are $\lambda_1, \dots, \lambda_k$ with

$$D_j g = \sum_{i=1}^k \lambda_i D_j f^i \quad \text{for } j = 1, \dots, n.$$

4. (a) For the ruled surface $f(s, t) = c(s) + t\delta(s)$, show that $m = \langle c', \delta \times \delta' \rangle$, and

$$EG - F^2 = \langle \delta, \delta \rangle \cdot \langle c' + t\delta', c' + t\delta' \rangle - \langle c' + t\delta', \delta \rangle^2.$$

(b) If θ is the angle between δ and $c' + t\delta'$, then

$$\langle \delta, \delta \rangle \cdot \langle c' + t\delta', c' + t\delta' \rangle \cdot \cos^2 \theta = \langle \delta, c' + t\delta' \rangle^2.$$

So $EG - F^2 = EG \sin^2 \theta = |(c' + t\delta') \times \delta|^2$.

(c) For the standard parameterization, show that

$$\sigma' \times \delta = \langle \sigma', \delta \times \delta' \rangle \cdot \delta',$$

and deduce the formula for K on page 148.

5. (a) Let $a, b \in \mathbb{R}^3$ and let $v, w \in \mathbb{R}^3$ be non-parallel vectors. If $a + t_0v$ and $b + t_1w$ are the points on the lines $\{a + tv\}$ and $\{b + tw\}$ which are closest to each other, then the line from $a + t_0v$ to $b + t_1w$ must be perpendicular to both v and w . Conclude that

$$t_0 = \frac{\langle w, w \rangle \cdot \langle a - b, v \rangle - \langle w, v \rangle \cdot \langle a - b, w \rangle}{\langle v, w \rangle^2 - \langle w, w \rangle \cdot \langle v, v \rangle}.$$

(b) Consider the ruled surface $c(s) + t\delta(s)$ with $|\delta| = 1$. If the point $P(\varepsilon)$ on page 147 is $c(s) + t(\varepsilon)\delta(s)$, then

$$t(\varepsilon) = \frac{\langle c(s) - c(s + \varepsilon), \delta(s) \rangle - \langle \delta(s), \delta(s + \varepsilon) \rangle \cdot \langle c(s) - c(s + \varepsilon), \delta(s + \varepsilon) \rangle}{\langle \delta(s), \delta(s + \varepsilon) \rangle^2 - 1}.$$

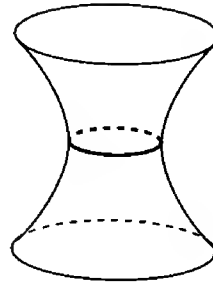
Use L'Hôpital's rule to show that

$$\lim_{\varepsilon \rightarrow 0} t(\varepsilon) = \frac{-\langle c'(s), \delta'(s) \rangle}{\langle \delta'(s), \delta'(s) \rangle}.$$

(c) The striction curve of the tangent developable of c is c .

(d) The striction curve of the hyperboloid of revolution

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} - \frac{z^2}{c^2} = 1$$



is the central circle. (Notice that in each case the tangent vector of the striction curve at a point is not perpendicular to the generator through that point, even though the striction curve is the limit of common perpendiculars to generators.)

6. (a) Modify Proposition II.4-14 as follows: If $\langle \cdot, \cdot \rangle$ is any (possibly degenerate) inner product on \mathbb{R}^n , then there is a basis of \mathbb{R}^n which is orthogonal for $\langle \cdot, \cdot \rangle$ and orthonormal with respect to the usual inner product.

(b) Consider a quadric $\{(x_1, x_2, x_3) : \sum_{ij} a_{ij}x_i x_j + \sum_i b_i x_i + c = 0\}$. Show that some rotation of \mathbb{R}^3 takes this into a set of the form

$$\{(x_1, x_2, x_3) : \sum_i \alpha_i x_i^2 + \sum_i \beta_i x_i + \gamma = 0\}.$$

(c) Show that a translation can be used to make $\beta_i = 0$ if $\alpha_i \neq 0$. Conclude that the quadric is an ellipsoid or hyperboloid of one or two sheets (or \emptyset), if all $\alpha_i \neq 0$; and it is an elliptic or hyperbolic paraboloid if just one $\alpha_i = 0$. Show that all other cases are lines, planes, or cylinders over parabolas, ellipses and hyperbolas.

7. Let $M \subset \mathbb{R}^3$ be a surface with unit normal $\nu: M \rightarrow \mathbb{R}^3$. We define the **support function** h of M by

$$h(p) = -\langle p, \nu(p) \rangle.$$

(a) Show that $|h(p)| = \text{distance from } 0 \text{ to } M_p$, and that $h(p) > 0$ if and only if $\nu(p)$ points toward the side of M_p which contains 0.

(b) For the ellipsoid $W^{-1}(0)$, where $W(x, y, z) = \frac{1}{2}(x^2/a^2 + y^2/b^2 + z^2/c^2 - 1)$, show that

$$h(x, y, z) = \frac{-1}{|Z|} \quad \text{for } Z = (W_1, W_2, W_3)(x, y, z).$$

Conclude that

$$K = \frac{h^4}{a^2 b^2 c^2},$$

and locate the points of maximum and minimum curvature.

(c) For the elliptic hyperboloids of one and two sheets, show that

$$K = \frac{-h^4}{a^2 b^2 c^2} \quad \text{and} \quad K = \frac{h^4}{a^2 b^2 c^2}, \quad \text{respectively.}$$

8. Let $M \subset \mathbb{R}^3$ be an imbedded surface such that $\nu: M \rightarrow S^2$ is one-one. For $\xi \in S^2$, let

$$p(\xi) = h(\nu^{-1}(\xi)),$$

where h is the support function of M .

(a) The tangent plane at $\nu^{-1}(\xi)$ is

$$\left\{ x \in \mathbb{R}^3 : \sum_{j=1}^3 \xi^j x^j = p(\xi) \right\},$$

so

$$\sum_{j=1}^3 \xi^j [\nu^{-1}(\xi)]^j = p(\xi).$$

(b) For $x \in \mathbb{R}^3 - \{0\}$, let

$$P(x) = |x| \cdot p\left(\frac{x}{|x|}\right).$$

Then

$$\frac{\partial P(x)}{\partial x^i} = \left[\nu^{-1}\left(\frac{x}{|x|}\right) \right]^i + \sum_{j=1}^3 x^j \cdot \frac{\partial \left[\nu^{-1}\left(\frac{x}{|x|}\right) \right]^j}{\partial x^i} = \left[\nu^{-1}\left(\frac{x}{|x|}\right) \right]^i.$$

Hint: The vanishing of the second term is equivalent to the assertion that $\partial[\nu^{-1}(x/|x|)]/\partial x^i$ is tangent to M at $\nu^{-1}(x/|x|)$. Note that $\nu^{-1}(x/|x|) \in M$ for all x .

9. (a) The determinant on page 152 equals

$$xa_2a_3(c^2 - b^2) + ya_1a_3(a^2 - c^2) + za_1a_2(b^2 - a^2).$$

By sign considerations, show that for $a > b > c > 0$ there is no umbilic with $y \neq 0$. Then show that there are four umbilics, with coordinates

$$x = \pm a \left(\frac{a^2 - b^2}{a^2 - c^2} \right)^{1/2} \quad z = \pm c \left(\frac{b^2 - c^2}{a^2 - c^2} \right)^{1/2}.$$

(b) Similarly, find the four umbilics on the elliptic hyperboloid of two sheets.

10. Find the umbilics on the elliptic paraboloid by using formulas (E'). [There are two if $a \neq b$, and one if $a = b$.]

11. (a) Let M be a doubly ruled non-flat surface, and choose three mutually skew straight lines L_1, L_2, L_3 from the first family of rulings. Show that there is a unique family of straight lines which intersect all three of L_1, L_2, L_3 .

(b) Show that three mutually skew lines L_1, L_2, L_3 lie on some quadric, and conclude that M is this quadric.

12. Let $M \subset \mathbb{R}^3$ be a surface with normal map ν . Then $\{p + \varepsilon\nu(p) : p \in M\}$ is called a **parallel surface** \bar{M} of M . We have a map $f: M \rightarrow \bar{M}$ given by $f(p) = p + \varepsilon\nu(p)$.

(a) If X is a tangent vector of M , then $f_*X = X + \varepsilon d\nu(X)$ [identifying tangent vectors with elements of \mathbb{R}^3 as usual]. Hence f is an immersion if $\varepsilon \neq 1/k_i$ for either principal curvature k_i at any point p of M . In particular, if M is compact, then \bar{M} is a surface for small enough ε . (One could also use Theorem I.9-20.)

(b) The normal $\bar{\nu}$ at $p + \varepsilon\nu(p)$ is just $\nu(p)$. (One could also use a generalization of Problem I.9-28.)

(c) The principal curvatures of \bar{M} are

$$\frac{k_i}{1 + \varepsilon k_i},$$

so the Gaussian and mean curvatures of \bar{M} are

$$\bar{K} = \frac{K}{1 + \varepsilon H + \varepsilon^2 K} \quad \bar{H} = \frac{H + 2\varepsilon K}{1 + \varepsilon H + \varepsilon^2 K}.$$

(d) If M has constant Gaussian curvature $K > 0$, then some parallel surface has constant mean curvature, and if M has constant mean curvature $H \neq 0$, then some parallel surface has constant Gaussian curvature (Bonnet).

(e) The volume element $d\bar{A}$ of \bar{M} is related to the volume element dA of M by

$$f^*(d\bar{A}) = (1 + 2\varepsilon H + \varepsilon^2 K) dA.$$

(f) If M has mean curvature $H = 0$, and M is not part of a plane, then the area of \bar{M} is *smaller* than the area of M (Steiner).

13. Let c be an arclength parameterized curve. Recall that the “rectifying plane” of c is spanned by the tangent \mathbf{t} and binomial \mathbf{b} . Suppose that the family of rectifying planes has an envelope M . Show that c is a geodesic on M . (This is the reason for the word rectifying—the curve c is “made straight” or “rectified” on M .)

CHAPTER 4

CURVES ON SURFACES

In classical surface theory, a great deal of emphasis was placed on special curves lying within a surface. In addition, several new invariants can be defined for a curve c on a surface, apart from the curvature κ and torsion τ which c has as a curve in \mathbb{R}^3 . The total corpus of accumulated results exhibits—to me at least—the unappealing weightiness of a massive treatise on the conic sections. So in this chapter we will give the definitions and then explore only a few of the pertinent results, concentrating on those which are of importance later on.

Let $M \subset \mathbb{R}^3$ be an oriented surface with corresponding unit normal field ν , and let c be an arclength parameterized curve in M . Then we can consider the normal and tangential components of $c''(s)$,

$$\perp c''(s) = \langle c''(s), \nu(c(s)) \rangle \cdot \nu(c(s)),$$

$$\top c''(s) = \frac{D}{ds} c'(s), \quad \text{by Corollary 1-2.}$$

The normal component $\perp c''(s)$ is sometimes called the **normal curvature vector** of c at s , and

$$(1) \quad \kappa_n(s) = \langle c''(s), \nu(c(s)) \rangle$$

is called the **normal curvature** of c at s ; it is the signed length of the normal curvature vector. As we mentioned in Chapter 1, the tangential component $\top c''(s)$ is called the **geodesic curvature vector**. Using the orientation of M we can assign a sign to the length of this vector. To do this, we first choose the unit vector $\mathbf{u}(s) \in M_{c(s)}$ perpendicular to $c'(s)$ for which $(c'(s), \mathbf{u}(s))$ is positively oriented, so that

$$c'(s) \times \mathbf{u}(s) = \nu(c(s)).$$

Then we note that

$$\langle c'(s), c'(s) \rangle = 1 \implies \left\langle \frac{D}{ds} c'(s), c'(s) \right\rangle = 0,$$

so that $\top c''(s) = D/ds(c'(s))$ must be a multiple of $\mathbf{u}(s)$. So we can define the **geodesic curvature** $\kappa_g(s)$ of c at s by

$$(2) \quad \begin{aligned} \top c''(s) &= \kappa_g(s) \cdot \mathbf{u}(s) \\ &= \kappa_g(s) \cdot \nu(c(s)) \times c'(s). \end{aligned}$$

Thus κ_g is the signed length of $\mathbf{T}c''$. Obviously we have

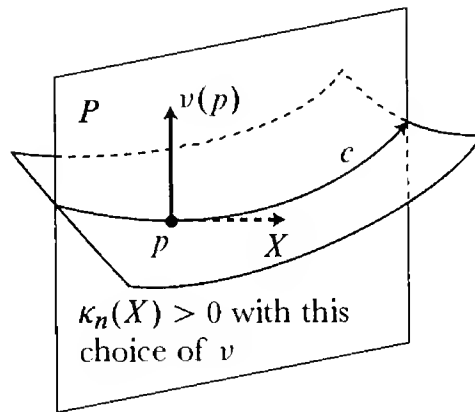
$$(3) \quad \kappa = \sqrt{\kappa_n^2 + \kappa_g^2}.$$

We will see that κ_n and κ_g have quite different properties.

We begin by recalling a few facts from Chapter II.3B. The equation $0 = \langle c'(s), \nu(c(s)) \rangle$ implies that

$$(4) \quad \begin{aligned} \kappa_n(s) &= \langle c''(s), \nu(c(s)) \rangle = - \left\langle \frac{d\nu(c(s))}{ds}, c'(s) \right\rangle = - \langle d\nu(c'(s)), c'(s) \rangle \\ &= \Pi(c'(s), c'(s)). \end{aligned}$$

Thus $\kappa_n(s)$ depends only on the direction $(c'(s))$ or $-c'(s)$ of c at s , and otherwise not on the curve c itself, so we can write $\kappa_n(X)$ for a unit vector X . Now for a given unit vector $X \in M_p$, there is a natural choice for a curve c in M with $c'(0) = X$, namely the arclength parameterized curve which is cut out on M by the plane P containing $\nu(p)$ and X . Then $\Pi(X, X) = \kappa_n(X)$ is

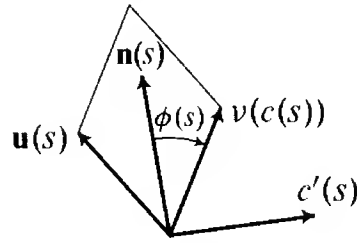


the signed curvature of this curve, when $(X, \nu(p))$ is chosen as the positive orientation for P . If $k_i = \kappa_n(X_i)$ are the minimum and maximum of these signed curvatures, so that X_i are eigenvectors of $-d\nu$, with eigenvalues k_i , then X_1 and X_2 are orthogonal, and if $X = (\cos \theta)X_1 + (\sin \theta)X_2$ is any unit vector, then (Euler's Theorem)

$$(5) \quad \kappa_n(X) = k_1 \cos^2 \theta + k_2 \sin^2 \theta.$$

Equation (4) can just as well be used to relate κ_n and κ for any curve c in M . Note that the normal $\mathbf{n}(s)$ of c at s is in the plane spanned by $\mathbf{u}(s)$ and $\nu(c(s))$.

Choosing $(\mathbf{u}(s), \nu(c(s)))$ as the positive orientation of this plane, we define $\phi(s)$



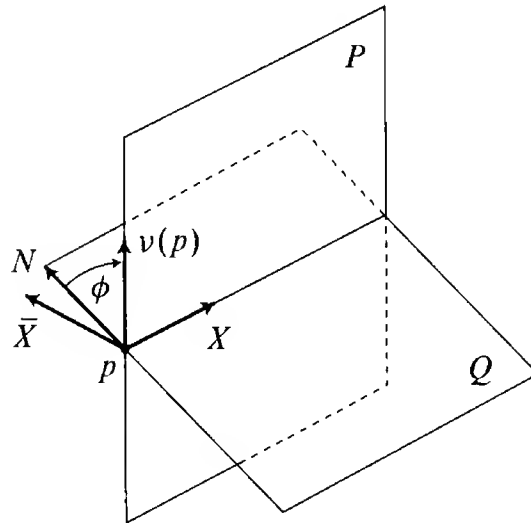
to be the oriented angle from $\mathbf{n}(s)$ to $\nu(c(s))$, so that

$$(6) \quad \mathbf{n}(s) = \sin \phi(s) \cdot \mathbf{u}(s) + \cos \phi(s) \cdot \nu(c(s)).$$

Then equation (4) implies that

$$(7) \quad \kappa_n(s) = \Pi(c'(s), c'(s)) = \kappa(s) \cdot \cos \phi(s),$$

where $\kappa (> 0)$ is the curvature of c . (If $\kappa(s) = 0$, then $\mathbf{n}(s)$ is undefined, so $\phi(s)$ is also; but equation (7) then holds with any choice of $\phi(s)$.) In particular, suppose that Q is any plane containing a unit vector $X \in M_p$. Let $\bar{X} \in M_p$ be the unit vector perpendicular to X for which (X, \bar{X}) is positively oriented, and



let $N \in \mathbb{R}_p^3$ be a unit vector in Q which is perpendicular to X . Choose (X, N) as the positive orientation for Q , and let ϕ be the oriented angle from N to $\nu(p)$ when $(\bar{X}, \nu(p))$ is chosen as the positive orientation for the plane perpendicular to X (changing N to $-N$ reverses the orientation of Q , and changes ϕ to $\phi - \pi$). If c_ϕ is the arclength parameterized curve cut out on M by Q , then its curvature κ_ϕ is given by

$$(7') \quad \kappa_\phi \cdot \cos \phi = \kappa(X).$$

Equation (7) or (7') is known as *Meusnier's Theorem*; another formulation of this theorem appears in Problem 1.

We can already state a result which is trivial, but which we will need to refer to later on.

1. PROPOSITION. Let c be an arclength parameterized curve in a surface $M \subset \mathbb{R}^3$, with $c(0) = p$, and $c'(0) = X \in M_p$.

(1) If X is an asymptotic vector, then either $\kappa(0) = 0$ or $\mathbf{n}(0)$ is perpendicular to $\nu(p)$.

(2) If X is not an asymptotic vector, then c cannot have curvature 0 at p , nor can $\mathbf{n}(0)$ be perpendicular to $\nu(p)$; if $\kappa_n(X) > 0$, then the angle between $\mathbf{n}(0)$ and $\nu(p)$ is acute, and if $\kappa_n(X) < 0$ it is obtuse (in the picture on page 188 the angle is 0).

PROOF. Everything follows from equation (4). ♦

Note, finally, that if c is not parameterized by arclength, then

$$(8) \quad \kappa_n(t) = \frac{\Pi(c'(t), c'(t))}{|c'(t)|^2},$$

since both numerator and denominator are multiplied by the same number under change of parameter.

Now consider the geodesic curvature $\kappa_g(s)$ of c at s , given by equation (2),

$$\mathbf{T}c''(s) = \kappa_g(s) \cdot \mathbf{u}(s).$$

Since we also have

$$c'' = \kappa \cdot \mathbf{n},$$

equation (6) immediately implies that

$$(9) \quad \kappa_g = \kappa \cdot \sin \phi.$$

(Again, this equation holds for any choice of ϕ when $\kappa = 0$.) Unlike κ_n , the quantity $\kappa_g = \text{signed length of } D/ds(c'(s))$ is *intrinsic*—it can be calculated directly from E, F, G using the formula on pg. II.232; one has to be careful to parameterize by arclength. On the other hand, $\kappa_g(s)$ does not depend only on c' . Indeed, κ_g is identically zero if c is a geodesic, and there are geodesics with arbitrary unit tangent vectors at any point.

Moving frame freaks will be happy to learn that κ_n and κ_g arise naturally when one chooses the appropriate moving frame along c . In fact, the original use of the moving frame was by Darboux, whose monumental 4 volume work on surfaces includes incredibly detailed investigations of curves on surfaces. Given

an arclength parameterized curve c in $M \subset \mathbb{R}^3$, we define the **Darboux frame** of c on M to be the moving frame

$$\mathbf{t}(s) = c'(s), \quad \mathbf{u}(s), \quad \mathbf{v}(s) = \mathbf{t}(s) \times \mathbf{u}(s) = \nu(c(s)),$$

as opposed to the **Frenet frame** $\mathbf{t}, \mathbf{n}, \mathbf{b}$. The Darboux frame is defined at all points of c , even those where $\kappa = 0$. Since the Darboux frame is also orthonormal, the expression for $(\mathbf{t}, \mathbf{u}, \mathbf{v})'$ is given by a skew-symmetric matrix times $(\mathbf{t}, \mathbf{u}, \mathbf{v})$:

$$(10) \quad \begin{aligned} \mathbf{t}' &= \kappa_g \mathbf{u} + \kappa_n \mathbf{v} \\ \mathbf{u}' &= -\kappa_g \mathbf{t} + \tau_g \mathbf{v} \\ \mathbf{v}' &= -\kappa_n \mathbf{t} - \tau_g \mathbf{u} \end{aligned}$$

From the moving frame point of view, the functions $\kappa_g, \kappa_n, \tau_g$ appearing here are defined by these equations, although it is clear from the first equation that κ_g and κ_n are the same as previously defined. The function τ_g is called the **geodesic torsion** of c . We proceed to indicate how these functions are analyzed using moving frames.

We first observe that $\mathbf{v}'(0)$ depends only on $\mathbf{t}(0)$, since

$$(11) \quad \mathbf{v}'(0) = \left. \frac{d\nu(c(s))}{ds} \right|_{s=0} = d\nu(\mathbf{t}(0)).$$

The third equation in (10) then shows immediately that κ_n and τ_g depend only on \mathbf{t} , so that we can write $\kappa_n(X)$ and $\tau_g(X)$ for unit vectors X . In fact, if $\bar{X} \in M_p$ is the unit vector perpendicular to X with (X, \bar{X}) positively oriented, then equations (10) and (11) give

$$(12) \quad \begin{aligned} \kappa_n(X) &= -\langle d\nu(X), X \rangle = \Pi(X, X) \\ \tau_g(X) &= -\langle d\nu(X), \bar{X} \rangle = \Pi(X, \bar{X}). \end{aligned}$$

Now let $X_1, X_2 \in M_p$ be principal directions with (X_1, X_2) positively oriented, and let k_1, k_2 be the corresponding principal curvatures. If θ is the oriented angle from X_1 to a unit vector $X \in M_p$, then we have

$$\begin{aligned} \kappa_n(X) &= -\langle d\nu(X), X \rangle \\ &= \langle k_1(\cos \theta)X_1 + k_2(\sin \theta)X_2, (\cos \theta)X_1 + (\sin \theta)X_2 \rangle \\ &= k_1 \cos^2 \theta + k_2 \sin^2 \theta. \end{aligned}$$

Here we have merely rederived Euler's Theorem. But exactly the same procedure gives an explicit expression for τ_g :

2. PROPOSITION. Let $X_1, X_2 \in M_p$ be principal directions, with (X_1, X_2) positively oriented, and let k_1, k_2 be the corresponding principal curvatures at p . If θ is the oriented angle from X_1 to a unit vector $X \in M_p$, then

$$\tau_g(X) = (k_2 - k_1) \sin \theta \cos \theta.$$

PROOF. Equation (12) gives

$$\begin{aligned} \tau_g(X) &= -\langle dv(X), \bar{X} \rangle \\ &= \langle k_1(\cos \theta)X_1 + k_2(\sin \theta)X_2, -(\sin \theta)X_1 + (\cos \theta)X_2 \rangle \\ &= (k_2 - k_1) \sin \theta \cos \theta. \quad \blacklozenge \end{aligned}$$

Now let ϕ be the oriented angle from \mathbf{n} to \mathbf{v} , as on page 189, so that

$$\begin{aligned} \mathbf{n} &= \sin \phi \cdot \mathbf{u} + \cos \phi \cdot \mathbf{v} \\ \mathbf{b} &= -\cos \phi \cdot \mathbf{u} + \sin \phi \cdot \mathbf{v}, \end{aligned}$$

and hence

$$\begin{aligned} \mathbf{u} &= \sin \phi \cdot \mathbf{n} - \cos \phi \cdot \mathbf{b} \\ \mathbf{v} &= \cos \phi \cdot \mathbf{n} + \sin \phi \cdot \mathbf{b}. \end{aligned}$$

Using the first equation in (10), and the Serret-Frenet formulas, we have

$$\begin{aligned} \kappa_g &= \langle \mathbf{u}, \mathbf{t}' \rangle = \langle (\sin \phi)\mathbf{n} - (\cos \phi)\mathbf{b}, \kappa \mathbf{n} \rangle \\ &= \kappa \cdot \sin \phi \\ \kappa_n &= \langle \mathbf{v}, \mathbf{t}' \rangle = \langle (\cos \phi)\mathbf{n} + (\sin \phi)\mathbf{b}, \kappa \mathbf{n} \rangle \\ &= \kappa \cdot \cos \phi, \end{aligned}$$

as before. (If $\kappa = 0$, then ϕ is undefined, but it follows immediately from the first equation of (10) that also $\kappa_g = \kappa_n = 0$.) We still have to make use of the second equation in (10); it will give us the geometric interpretation of τ_g . Now we have

$$\tau_g = \langle \mathbf{v}, \mathbf{u}' \rangle = \left\langle (\cos \phi)\mathbf{n} + (\sin \phi)\mathbf{b}, \frac{d}{ds}[(\sin \phi)\mathbf{n} - (\cos \phi)\mathbf{b}] \right\rangle;$$

using the Serret-Frenet formulas

$$\frac{d\mathbf{n}}{ds} = \mathbf{n}' = -\kappa \mathbf{t} + \tau \mathbf{b} \quad \text{and} \quad \frac{d\mathbf{b}}{ds} = -\tau \mathbf{n},$$

we end up with

$$(13) \quad \tau_g = \tau + \frac{d\phi}{ds}, \quad \text{whenever } \tau \text{ is defined.}$$

In particular we have

3. PROPOSITION.

- (a) If $X \in M_p$ is a unit vector, then $\tau_g(X)$ is the torsion $\tau(0)$ of the geodesic γ with $\gamma'(0) = X$.
- (b) The geodesics pointing in a principal direction at a point have torsion 0 at that point; in particular, all geodesics have torsion 0 at an umbilic.
- (c) If $X, Y \in M_p$ are perpendicular unit vectors, then $\tau_g(X) = -\tau_g(Y)$; thus orthogonal geodesics through a point have torsions at that point which are negatives of each other.

Remark: These statements hold only with the additional *proviso* that the torsions in question exist. Problem 3 considers what happens otherwise.

PROOF. Part (a) follows from equation (13), since for the geodesic γ we have $\phi = 0$ or $\phi = \pi$ for all s . Then parts (b) and (c) follow from Proposition 2. ♦

We note in passing that changing the direction of a curve c in M changes \mathbf{t} to $-\mathbf{t}$, and \mathbf{u} to $-\mathbf{u}$, but leaves \mathbf{v} fixed. So equations (10) show that κ_g changes sign, while κ_n and τ_g remain the same. For κ_n this follows from equation (7), since \mathbf{n} is changed to $-\mathbf{n}$, so ϕ is changed to $\phi - \pi$. It also follows from the interpretation of $\kappa_n(X)$ as the signed curvature of the curve cut out on M by the plane P through X and $\nu(p)$ —normally, reversing the curve would change the curvature, but in this case, since we change X to $-X$, we also change the orientation of P . The fact that τ_g remains the same follows from the fact that reversing the direction of a curve does not change its torsion.

One other fact about our new invariants will be of interest. An old theorem of Laguerre says that, like κ_n and τ_g , the quantity

$$\frac{d\kappa_n(s)}{ds} - 2\tau_g(s)\kappa_g(s)$$

also depends only on $c'(s)$; using equations (7), (9), and (13), this quantity can be written in Laguerre's formulation, involving ϕ , κ , and τ , as

$$\frac{d\kappa(s)}{ds} \cos \phi(s) - \left(3 \frac{d\phi(s)}{ds} + 2\tau(s) \right) \kappa(s) \sin \phi(s),$$

which makes the result seem even more mysterious. Élie Cartan observed that just as κ_n and τ_g can be expressed in terms of the tensor \mathbf{II} , this new expression,

and others like it, can be expressed in terms of the *covariant derivatives* of Π . Recall that by Corollary II.6-5 the tensor $(\nabla_{Z_p}\Pi)(X_p, Y_p) = (\nabla\Pi)(X_p, Y_p, Z_p)$ satisfies

$$(\nabla_{Z_p}\Pi)(X_p, Y_p) = Z_p(\Pi(X, Y)) - \Pi(\nabla_{Z_p}X, Y_p) - \Pi(X_p, \nabla_{Z_p}Y),$$

for all vector fields X, Y, Z extending X_p, Y_p, Z_p . This immediately yields

4. PROPOSITION. For all arclength parameterized curves c in $M \subset \mathbb{R}^3$ with the same tangent vector $c'(0) \in M_p$, the quantity

$$\kappa_n'(s) - 2\tau_g(s)\kappa_g(s)$$

has the same value at $s = 0$. The same is true for

$$\tau_g'(s) + 2[\kappa_n(s) - H(c(s))]\kappa_g(s).$$

PROOF. Let X be a unit vector field on M which extends \mathbf{t} , and let \bar{X} be the perpendicular unit vector field with (X, \bar{X}) positively oriented. Then equation (2), and the fact that $\mathbf{T}c''(s) = D/ds(c'(s))$, shows that

$$\nabla_X X = \kappa_g \cdot \bar{X} \quad \text{along } c,$$

so by equation (12) we have

$$\Pi(\nabla_X X, X) = \kappa_g \cdot \tau_g \quad \text{along } c.$$

Hence

$$\begin{aligned} (\nabla_X \Pi)(X, X) &= X(\Pi(X, X)) - 2\Pi(\nabla_X X, X) \\ &= \kappa_n' - 2\kappa_g \tau_g \quad \text{along } c. \end{aligned}$$

This shows that the first expression depends only on $X = c'$.

We also have

$$\begin{aligned} \langle X, \bar{X} \rangle &= 0 \implies \langle \nabla_X \bar{X}, X \rangle = -\langle \bar{X}, \nabla_X X \rangle = -\kappa_g, \\ \langle \bar{X}, \bar{X} \rangle &= 1 \implies \langle \nabla_X \bar{X}, \bar{X} \rangle = 0, \end{aligned}$$

which implies that

$$\nabla_X \bar{X} = -\kappa_g X \quad \text{along } c.$$

So we obtain

$$\begin{aligned}
 (\nabla_X \Pi)(X, \bar{X}) &= X(\Pi(X, \bar{X})) - \Pi(\nabla_X X, \bar{X}) - \Pi(X, \nabla_X \bar{X}) \\
 &= \tau_g' - \kappa_g \Pi(\bar{X}, \bar{X}) + \kappa_g \Pi(X, X) \\
 &= \tau_g' + \{2\Pi(X, X) - [\Pi(X, X) + \Pi(\bar{X}, \bar{X})]\} \kappa_g \\
 &= \tau_g' + 2(\kappa_n - H)\kappa_g \quad \text{along } c. \quad \spadesuit
 \end{aligned}$$

We are now ready to consider the three main classes of curves on a surface $M \subset \mathbb{R}^3$. The curve c is called a **line of curvature** (or **principal curve**) if c' always points along a principal direction. This means that

$$-dv(c') = k \cdot c'$$

for some function k , where $k(t)$ must be a principal curvature at $c(t)$. We can also write this as

$$\frac{-dv(c(t))}{dt} = k(t) \frac{dc}{dt} \quad \text{or} \quad \frac{dv}{dt} + k \frac{dc}{dt} = 0, \quad \text{in the classical manner.}$$

Oddly enough, the last equation has a special name, **Rodrigues' formula**. A more interesting characterization of principal curves can be given in terms of one of our invariants—the third equation in (10), together with (11), shows that c is a line of curvature if and only if τ_g is identically zero.

A curve c in M is called an **asymptotic curve** if c' always points along an asymptotic direction. Equation (4) [or (8)] shows that c is an asymptotic curve if and only if $\mathbf{n}(t)$ is perpendicular to $v(c(t))$ at all points t where $\kappa(t) \neq 0$. Equivalently, c is an asymptotic curve if and only if $\mathbf{n}(t)$ lies in $M_{c(t)}$ whenever $\kappa(t) \neq 0$, or yet again, if and only if the osculating plane of c at t coincides with $M_{c(t)}$ whenever $\kappa(t) \neq 0$. Equation (4), or the third equation in (10), together with (11), shows that c is an asymptotic curve if and only if κ_n is identically zero. Equation (3), or the first equation of (10), then shows that c is an asymptotic curve if and only if $\kappa = \pm \kappa_g$ everywhere. Moreover, at points where $\kappa \neq 0$, we then have $\mathbf{n} = \pm \mathbf{u}$, so $\mathbf{b} = \pm \mathbf{v}$ (the same sign holding in both cases), and the third equation of (10) shows that $\tau = \tau_g$; this also follows from equation (13), since $\phi = \pi/2$ or $3\pi/2$ for all t .

Finally, we have the geodesics, which may be defined as curves c with $c''(t)$ always perpendicular to $M_{c(t)}$, so that \mathbf{n} is perpendicular to M at points where $\kappa \neq 0$, or as curves with κ_g identically zero.

To summarize, we have

$$\begin{aligned}
 c \text{ is a line of curvature} &\iff -dv(c') = k \cdot c' \\
 &\iff \tau_g = 0 \\
 c \text{ is an asymptotic curve} &\iff \mathbf{n}, \text{ when defined, is} \\
 &\quad \text{always tangent to } M \\
 &\iff \text{the osculating plane, when defined,} \\
 &\quad \text{always coincides with the} \\
 &\quad \text{tangent plane of } M \\
 &\iff \kappa_n = 0 \\
 &\iff \kappa = \pm \kappa_g \\
 &\implies \tau = \tau_g, \text{ when } \tau \text{ is defined} \\
 c \text{ is a geodesic} &\iff \mathbf{n}, \text{ when defined, is} \\
 &\quad \text{always perpendicular to } M \\
 &\iff \kappa_g = 0 \\
 &\implies \tau = \tau_g, \text{ when } \tau \text{ is defined.}
 \end{aligned}$$

Since κ_n and τ_g actually depend only on the direction of c at a point, it also makes sense to talk about a curve c being “asymptotic at t ” ($\kappa_n(t) = 0$) or “principal at t ” ($\tau_g(t) = 0$); this just means that $c'(t)$ points in an asymptotic direction or in a principal direction. The equivalences given above for lines of curvature and asymptotic curves can all be replaced by corresponding equivalences for curves which are principal at a point or asymptotic at a point; however the conclusion $\tau(t) = \tau_g(t)$ does not follow from the mere assumption that c is asymptotic at t .

We will begin our study of these special curves by considering some very general properties. Taking the asymptotic curves first, we note that they can exist only in regions where $K \leq 0$. This already leads to another simple

5. PROPOSITION. A straight line on a surface is an asymptotic curve, so the curvature K of the surface satisfies $K \leq 0$ along any straight line lying in it. The curvature is everywhere 0 along the straight line if and only if the normal ν is constant along the line (equivalently: if and only if the tangent space is parallel along the line).

PROOF. The first assertion follows immediately from equation (4). To prove the second, let c be any parameterization of the straight line (with $c'(t)$ always

$\neq 0$). If the normal is constant along c , then

$$0 = \frac{dv(c(t))}{dt} = dv(c'(t)),$$

and this can happen only if dv has determinant $K = 0$. Conversely, suppose $K = 0$ at each point $c(t)$, so that dv has at least one eigenvalue 0 at $c(t)$. If both eigenvalues are 0, then certainly $dv(c'(t)) = 0$. On the other hand, it is easy to see that if one eigenvalue is non-zero, then the only asymptotic vectors are multiples of the eigenvector with eigenvalue 0; in other words, the asymptotic vector $c'(t)$ must satisfy $dv(c'(t)) = 0$. ♦

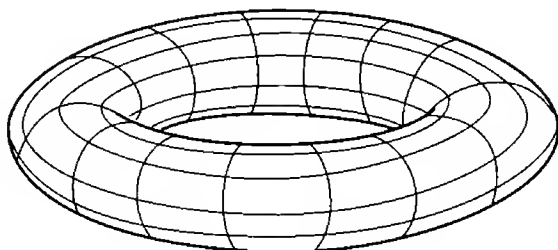
Remark: Actually, we have proved a more general result a long time ago (Corollary 1-7), but I thought it would be nice to include the classical proof also. As before, we can state a slightly more precise result: $K = 0$ at a point of a straight line if and only if at this point the normal v has derivative 0 along the line.

As an immediate corollary of Proposition 5, notice that a ruled surface must have $K \leq 0$ everywhere, as we found by computation in Chapter 3. We also see that $K = 0$ everywhere on a ruled surface if and only if the normal v is constant on each generator. On page 146 we found that the tangent plane at $f(s, t)$ is spanned by $c'(s) + t\delta'(s)$ and $\delta(s)$. This is independent of t if and only if $\delta(s), \delta'(s), c'(s)$ are linearly dependent. Our formula for K on page 147 shows that this is indeed true if and only if $K = 0$ everywhere. Classically, the ruled surfaces with $K = 0$ everywhere, i.e., the ruled surfaces with constant normals along each generator, were called **developable surfaces**; in the next chapter we will consider them in greater detail.

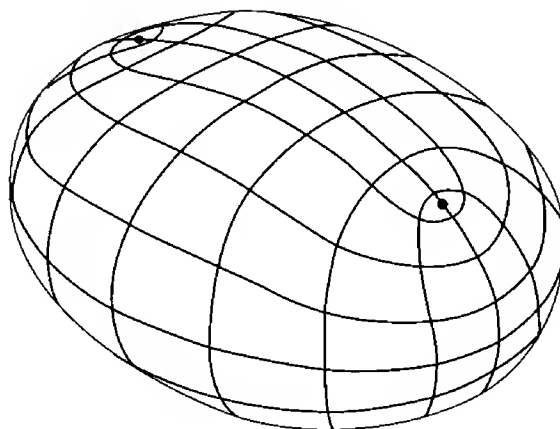
In a small region of a surface where $K < 0$, we can choose two linearly independent asymptotic unit vectors X_p, Y_p at each point p . It is easy to show that $p \mapsto X_p$ and $p \mapsto Y_p$ will be C^∞ vector fields; the asymptotic curves are just the integral curves of these C^∞ vector fields—as one approaches a parabolic or planar point these integral curves can run together in complicated ways. In Chapter 2 we pointed out that the asymptotic directions are perpendicular precisely when the mean curvature $H = 0$. So on a minimal surface without planar points, the asymptotic curves are everywhere orthogonal. This is illustrated on the right helicoid by the rulings and the helices.

In contrast to the asymptotic curves, lines of curvature can exist in regions of any sort, and it is only umbilics which cause problems. In a small region free of umbilics, we can choose two linearly independent principal unit vectors X_p, Y_p at each point p . As in the case of asymptotic directions, it is easy to show that $p \mapsto X_p$ and $p \mapsto Y_p$ are C^∞ vector fields (on regions where $K < 0$ this

is also a consequence of the fact that the principal directions bisect the asymptotic directions); the lines of curvature are the integral curves of these vector fields. We have already seen that on a torus of revolution, with no umbilics, the

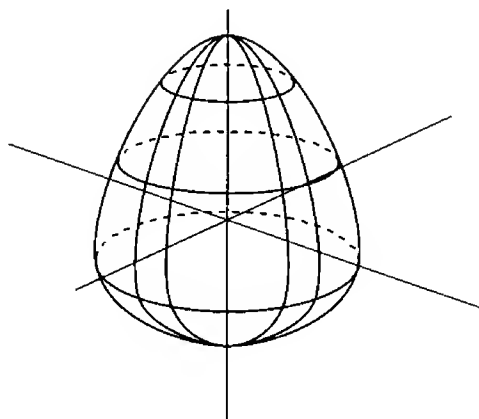


lines of curvature are the parallels and meridians. By contrast, the following famous picture shows how the lines of curvature behave in a neighborhood of the umbilic points of an ellipsoid.



Notice that if a surface M has no umbilics, like the torus, then there is a C^∞ 1-dimensional distribution on M —at each point p we choose the set of all vectors in M_p which are eigenvalues for the larger principal curvature, say. Equivalently, we can pick out two units vectors $X_p, -X_p \in M_p$. Now we can construct a 2-fold covering space $\pi: \tilde{M} \rightarrow M$ by choosing the two points in $\pi^{-1}(p)$ to correspond to these two vectors. There is then an obvious nowhere zero vector field \tilde{X} on \tilde{M} (if $q \in \pi^{-1}(p)$ is the point corresponding to $X_p \in M_p$, then $\pi_* \tilde{X}_q = X_p$). On the other hand, Theorem I.11-30 tells us that for the compact orientable surface \tilde{M} , this can happen only when the Euler characteristic $\chi(\tilde{M}) = 0$, which implies that $\chi(M) = 0$, since $\chi(\tilde{M}) = 2\chi(M)$ (Problem 7). But the torus is the *only* compact orientable surface with Euler characteristic $= 0$ (Problem I.11-2(c)). So *any compact surface in \mathbb{R}^3 not homeomorphic to the torus must have at least one umbilic*.

Carathéodory conjectured that every compact convex surface in \mathbb{R}^3 must have at least *two* umbilics. Weird as this may seem, there is a natural way to try to prove it. On any compact surface $M \subset \mathbb{R}^3$ with only finitely many umbilics p_1, \dots, p_k we can choose a C^∞ 1-dimensional distribution on $M - \{p_1, \dots, p_k\}$. There is a way of defining the index of this distribution at each point p_i , in much the same way that we defined the index of a vector field in Chapter I.11, except that now the index can take on half-integer values. Moreover, it turns out that the sum of the indices of the distribution is again the Euler characteristic $\chi(M)$. The precise definition of the index, and the proof of this result are given in Addendum 2. Now $\chi(S^2) = 2$, so Carathéodory's conjecture could be proved by showing that at an umbilic the index of our particular distribution cannot be equal to 2. For the analytic case, Hamburger [1] gave a proof of this which is 183 pages long! Bol [1] then gave a proof that is only 22 pages long, although it requires a correction (Klotz [1]). After all this work, it still seems that nothing is known when the surface is not analytic, or when it is not convex, even if it is homeomorphic to S^2 . I also know of no example where there are only two umbilics. On compact surfaces of revolution the lines of curvature have only two singularities, at the two poles, but I suspect that there will always be at least one whole parallel of umbilics in addition.



Finally, we have the geodesics. They, of course, not only exist in any sort of region, but can be found in any direction. On the other hand, just as the asymptotic lines intersect orthogonally only when $H = 0$, orthogonality of two families of geodesics implies that $K = 0$; in fact, even more is true:

6. PROPOSITION. If two families of geodesics intersect at a constant angle everywhere on M , then M is flat.

PROOF. Let X [or Y] be the vector field of unit tangents to the curves of the first [or second] family. Then X is parallel along the integral curves of X .

Since the angle between X and Y is constant, and since we are on a surface, the vector field Y must also be parallel along the integral curves of X . The same argument holds with X and Y interchanged. We therefore have

$$0 = \nabla_X X = \nabla_Y Y = \nabla_X Y = \nabla_Y X = [X, Y].$$

Consequently,

$$R(X, Y)Y = \nabla_X(\nabla_Y Y) - \nabla_Y(\nabla_X Y) - \nabla_{[X, Y]}Y = 0. \quad \blacklozenge$$

Proposition 6, of course, is not really a theorem about surfaces in \mathbb{R}^3 at all—it is actually a theorem about the intrinsic geometry of surfaces (I do not know whether any analogue holds for higher dimensional manifolds).

With this very general discussion of the behavior of our three classes of curves out of the way, we proceed to the main results about each class. For asymptotic curves this result is

7. THEOREM (BELTRAMI-ENNEPER). If c is an asymptotic curve with $c(0) = p$, and $\kappa(0) \neq 0$, then

$$|\tau(0)| = \sqrt{-K(p)}.$$

Moreover, if $K(p) < 0$ and the two distinct asymptotic curves through p both have non-zero curvature at p , then their torsions at p are negatives of each other.

FIRST (SEMI-) PROOF. Parameterize c by arclength. We know that both \mathbf{t} and \mathbf{n} lie in the tangent space of M , and we can let $\mathbf{v} = \mathbf{b} = \mathbf{t} \times \mathbf{n}$. Then the Serret-Frenet formulas give

$$-d\mathbf{v}(\mathbf{t}(s)) = -\mathbf{b}'(s) = \tau(s) \cdot \mathbf{n}(s).$$

So the matrix of the self-adjoint transformation $-d\mathbf{v}: M_p \rightarrow M_p$, with respect to the orthonormal basis $\mathbf{t}(0), \mathbf{n}(0)$ must be the symmetric matrix

$$\begin{pmatrix} 0 & \tau(0) \\ \tau(0) & 0 \end{pmatrix},$$

with determinant $K(p) = -\tau(0)^2$. This proof does not give any information about the sign of τ .

SECOND PROOF. We know that $\tau = \tau_g$ for the asymptotic curve c . So Proposition 2 gives

$$(1) \quad \tau(0) = (k_2 - k_1) \sin \theta \cos \theta,$$

where θ is the oriented angle from the principal vector X_1 to $X = c'(0)$. On the other hand, since X is an asymptotic vector, Euler's formula [equation (5)] gives

$$(2) \quad 0 = k_1 \cos^2 \theta + k_2 \sin^2 \theta,$$

so

$$(3) \quad k_2 = -k_1 \frac{\cos^2 \theta}{\sin^2 \theta}.$$

Substituting into (1) we have

$$\begin{aligned} \tau(0) &= -k_1 \left(\frac{\cos^2 \theta}{\sin^2 \theta} + 1 \right) \sin \theta \cos \theta \\ &= -k_1 \cdot \frac{\cos \theta}{\sin \theta}, \end{aligned}$$

while

$$K(p) = k_1 k_2 = -k_1^2 \frac{\cos^2 \theta}{\sin^2 \theta} \quad \text{by (3).}$$

Hence $\tau(0)^2 = -K(p)$ [we divided by $\sin \theta$, but if $\sin \theta = 0$, we could instead solve for k_1 in terms of k_2 ; alternatively, we can simply note that if $\sin \theta = 0$, then we have $k_1 = 0$ from (2) and $\tau(0) = 0$ from (1)].

Equation (1) also shows that $\tau(0)$ changes sign when we change θ to $-\theta$, which gives the second part of the theorem. ♦

As an example, on the right helicoid we can choose as one family of asymptotic curves the helices $c(s) = (t \cos s, t \sin s, bs)$, for fixed t . On pg. II.33 we found that

$$\tau = \frac{b}{b^2 + t^2},$$

in agreement with our formula for K (page 150). In this example the other family of asymptotic curves are straight lines, with vanishing curvature.

The Beltrami-Enneper theorem raises a natural question, one so natural that no one since Darboux seems to have considered it. Given a unit vector $X \in M_p$,

we can consider the line of curvature, the asymptotic line, and the geodesic which have tangent vector X at p (for simplicity we assume p is not an umbilic for the case of lines of curvature, and $K(p) < 0$ for the case of an asymptotic curve). We know that

- the torsion of the asymptotic curve c with $c'(0) = X$ is $\pm\sqrt{-K(p)}$,
- the torsion of the geodesic c with $c'(0) = X$ is $\tau_g(X) = \Pi(X, \bar{X})$,
- the curvature of the geodesic c with $c'(0) = X$ is $|\kappa_n(X)| = |\Pi(X, X)|$;

the first statement is Theorem 7, the second is Proposition 3, and the third follows from the equation $\kappa_n^2 + \kappa_g^2 = \kappa^2$, since $\kappa_g = 0$ for a geodesic. Now it is just as reasonable to ask for the curvature of the asymptotic curve c with $c'(0) = X$. To determine it, we can use one of the invariants of Proposition 4. We saw, in the proof of that Proposition, that

$$(\nabla_X \Pi)(X, X) = \kappa_n'(0) - 2\tau_g(0)\kappa_g(0)$$

for *any* curve c with $c'(0) = X$. Now if c is an asymptotic curve, then κ_n is identically zero, while $\kappa = \pm\kappa_g$ and $\tau = \tau_g$, so we have

$$\begin{aligned} (\nabla_X \Pi)(X, X) &= 0 \mp 2\tau(0)\kappa(0) \\ &= \mp 2\sqrt{-K(p)}\kappa(0), \quad \text{by Theorem 7.} \end{aligned}$$

Hence, if $K(p) \neq 0$, then

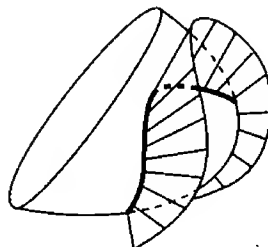
the curvature of the asymptotic curve c with $c'(0) = X$ is

$$\frac{|(\nabla_X \Pi)(X, X)|}{2\sqrt{-K(p)}}.$$

We can also ask for the curvature and torsion of a line of curvature c . Here the situation is somewhat different, since $c'(0) = X \in M_p$ is already essentially determined. We have $\tau_g = 0$ for lines of curvature, so the second invariant of Proposition 4 gives us the value of $[\kappa_n(X) - H(p)]\kappa_g(0)$, and hence of $\kappa_g(0)$. Then we can determine $\kappa = \sqrt{\kappa_g^2 + \kappa_n^2}$ at 0. To compute the torsion $\tau(0)$, we have to use yet another invariant, involving *second* derivatives (Problems 5 and 6).

To study lines of curvature, we begin with a pretty, though not very useful, criterion for such curves.

8. **THEOREM (BONNET).** A curve c in M is a line of curvature if and only if the surface S formed by the normals to the surface along c is flat.



PROOF. The surface S is the ruled surface parameterized by

$$f(s, t) = c(s) + tv(c(s)) = c(s) + t\delta(s), \quad \text{say.}$$

From page 147 we see that S is flat if and only if

$$0 = \langle c'(s), \delta(s) \times \delta'(s) \rangle = \left\langle c'(s), v(c(s)) \times \frac{dv(c(s))}{ds} \right\rangle,$$

which is true if and only if $dv(c(s))/ds$ is a multiple of $c'(s)$. ♦

In the case of a surface of revolution, Theorem 8 shows that meridians and parallels must be lines of curvature, since the corresponding surfaces S are planes and cones. (Of course, we have already argued in essentially just this way on pages 158–159.) The next theorem can also be used to find the lines of curvature on a surface of revolution.

9. **THEOREM (TERQUEM-JOACHIMSTHAL).** Let c be a curve in $M_1 \cap M_2$ which is a line of curvature in M_1 . Then c is a line of curvature in M_2 if and only if M_1 and M_2 intersect at a constant angle along c (i.e., the normals of M_1 and M_2 have the same angle along c).

PROOF. If v_i is the unit normal field on M_i , then

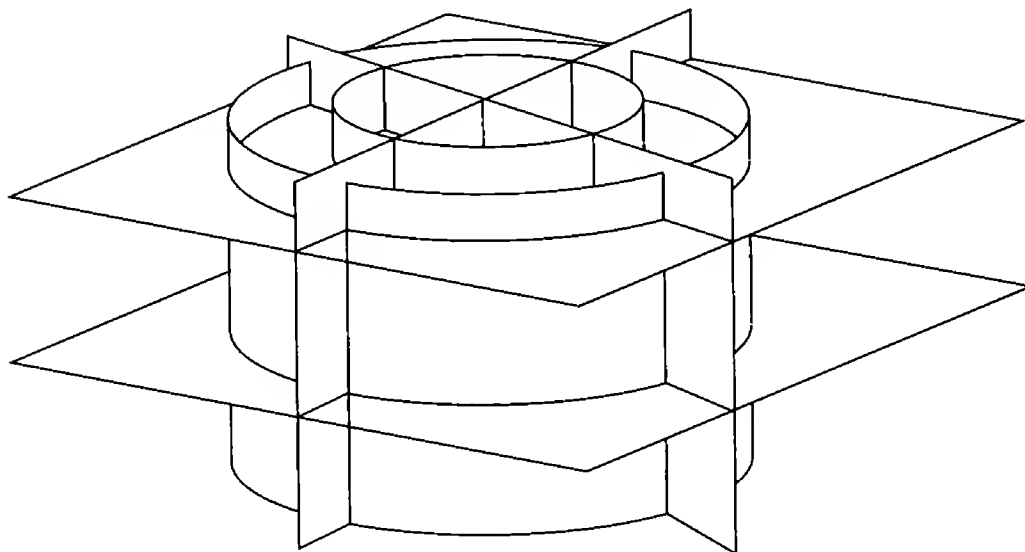
$$\begin{aligned} \frac{d}{ds} \langle v_1(c(s)), v_2(c(s)) \rangle &= \left\langle \frac{dv_1(c(s))}{ds}, v_2(c(s)) \right\rangle + \left\langle v_1(c(s)), \frac{dv_2(c(s))}{ds} \right\rangle \\ &= \left\langle -k(s) \frac{dc}{ds}, v_2(c(s)) \right\rangle + \left\langle v_1(c(s)), \frac{dv_2(c(s))}{ds} \right\rangle, \\ &\quad \text{since } c \text{ is a line of curvature of } M_1 \\ &= 0 + \left\langle v_1(c(s)), \frac{dv_2(c(s))}{ds} \right\rangle, \end{aligned}$$

since c is a curve in M_2 . If c is a line of curvature in M_2 , then the remaining term is similarly 0, so $\langle v_1(c(s)), v_2(c(s)) \rangle$ is constant.

Conversely, if this quantity has derivative 0, then $dv_2(c(s))/ds$ is perpendicular to $v_1(c(s))$. On the other hand, it is also perpendicular to $v_2(c(s))$. If $v_1(c(s))$ and $v_2(c(s))$ are linearly independent, then $dv_2(c(s))/ds$ must be a multiple of $c'(s)$, and consequently c is a line of curvature in M_2 . If $v_1(c(s))$ and $v_2(c(s))$ are *not* linearly independent, then we must have $v_1(c(s)) = \pm v_2(c(s))$ for all s (since $\langle v_1(c(s)), v_2(c(s)) \rangle$ is constant). In this case there is nothing left to prove. ♦

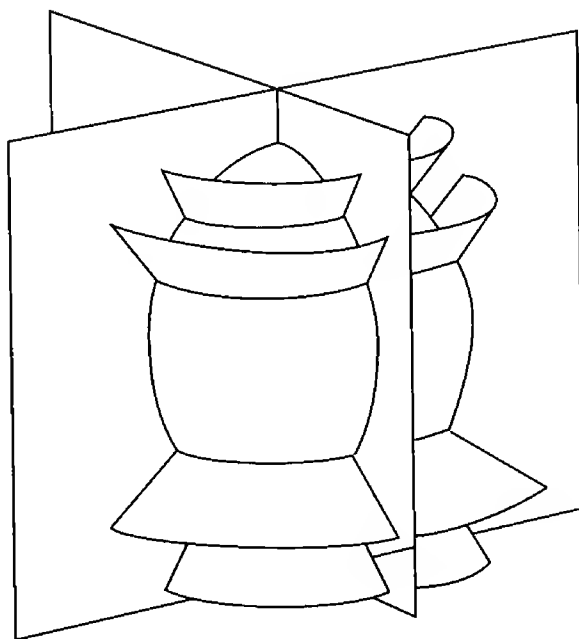
The one really interesting result about lines of curvature concerns **triply orthogonal systems** of surfaces—these are triples of 1-parameter families of surfaces with the property that at each point the tangent planes of the surfaces from any two families are perpendicular. The simplest examples of triply orthogonal systems are the following:

- (1) Each family consists of all the planes that are parallel to one of the coordinate planes.
- (2) The first family consists of all planes parallel to the (x, y) -plane; the second family consists of all the circular cylinders having the z -axis as their common axis; the third family consists of all planes that pass through the z -axis.



- (3) The first family consists of all the concentric spheres around the point 0; the second family consists of all planes that pass through the z -axis; the

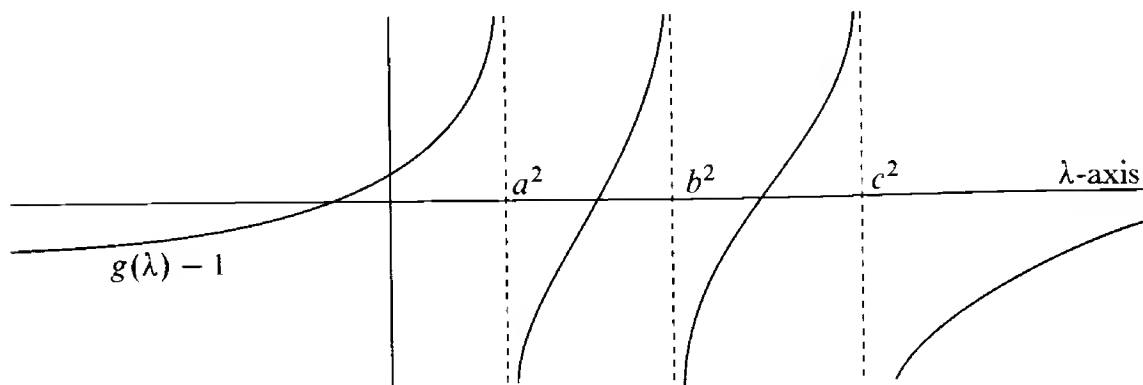
third family consists of cones, each cone being formed by the all the lines through 0 that make some fixed angle with the z -axis.



The one other, less trivial, standard example is formed by the set of all surfaces satisfying the equation

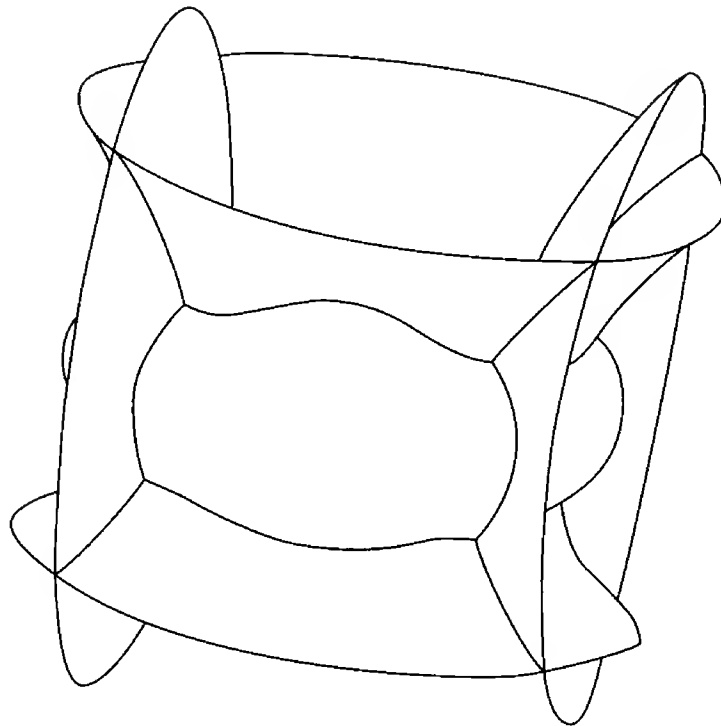
$$(*) \quad g(\lambda) = \frac{x^2}{a^2 - \lambda} + \frac{y^2}{b^2 - \lambda} + \frac{z^2}{c^2 - \lambda} = 1, \quad 0 < a^2 < b^2 < c^2.$$

For $\lambda < a^2$ we obtain ellipsoids, for $a^2 < \lambda < b^2$ hyperboloids of one sheet, and for $b^2 < \lambda < c^2$ hyperboloids of two sheets. For any (x, y, z) with $x, y, z \neq 0$, the function $\lambda \mapsto g(\lambda) - 1$ is continuous except at a^2, b^2, c^2 ; it clearly jumps from $+\infty$ to $-\infty$ as we pass from the left of one of these points to the right of it, and $g(\lambda) - 1 \rightarrow -1$ as $\lambda \rightarrow -\infty$. Consequently, $g(\lambda) - 1$ must be 0 for at



least one $\lambda_1 < a^2$, one λ_2 with $a^2 < \lambda_2 < b^2$, and one λ_3 with $b^2 < \lambda_3 < c^2$.

There are only 3 roots, since $g(\lambda) - 1 = 0$ is equivalent to a cubic equation in λ . Thus one surface from each family passes through each such point (x, y, z) . At



a point (x, y, z) on the surface $g(\lambda_i) = 1$, the normal vector has the direction

$$\frac{1}{2}(D_1 g(\lambda_i), D_2 g(\lambda_i), D_3 g(\lambda_i)) = \left(\frac{x}{a^2 - \lambda_i}, \frac{y}{b^2 - \lambda_i}, \frac{z}{c^2 - \lambda_i} \right).$$

At a point (x, y, z) on the two surfaces $g(\lambda_i) = 1$ and $g(\lambda_j) = 1$, the inner product of the two normal vectors is therefore

$$\frac{x^2}{(a^2 - \lambda_i)(a^2 - \lambda_j)} + \frac{y^2}{(b^2 - \lambda_i)(b^2 - \lambda_j)} + \frac{z^2}{(c^2 - \lambda_i)(c^2 - \lambda_j)}$$

which can be written as

$$\frac{g(\lambda_i) - g(\lambda_j)}{\lambda_j - \lambda_i} = 0.$$

Thus our system is orthogonal. Since we can always imbed a given ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ in a system of the form $(*)$, the following result enables us to describe the lines of curvature on an ellipsoid.

10. THEOREM (DUPIN). The lines of intersection of the surfaces of a triply orthogonal system are lines of curvature on the surfaces.

PROOF. Let Δ_i be the distribution formed by the tangent planes to the i^{th} family of surfaces. Pick unit vector fields X, Y, Z with

$$X \in \Delta_1 \cap \Delta_3, \quad Y \in \Delta_2 \cap \Delta_3, \quad Z \in \Delta_1 \cap \Delta_2.$$

Letting ∇' be the ordinary directional derivative in \mathbb{R}^3 , we have

$$\nabla'_X Y - \nabla'_Y X = [X, Y] \in \Delta_3, \quad \text{since } \Delta_3 \text{ is integrable (pg. I.192).}$$

Using orthogonality, we conclude that

$$(1) \quad \langle \nabla'_X Y, Z \rangle = \langle \nabla'_Y X, Z \rangle,$$

and of course we can permute X, Y, Z in this equation. On the other hand, we also have

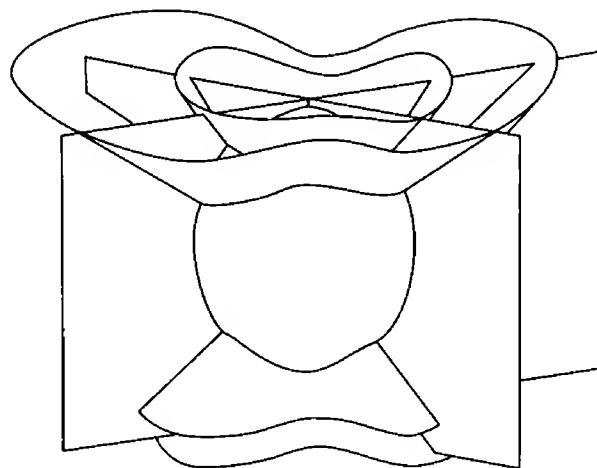
$$(2) \quad 0 = X(\langle Y, Z \rangle) = \langle \nabla'_X Y, Z \rangle + \langle Y, \nabla'_X Z \rangle,$$

together with the equations obtained by permuting X, Y, Z . From the equations comprised in (1) and (2) we can conclude, for example, that

$$0 = \langle \nabla'_X Y, Z \rangle = \langle \nabla'_X Y, Y \rangle,$$

so that $\nabla'_X Y$ must be a multiple of X . It follows that the line of intersection of two surfaces in the first and third family is a line of curvature on the surface in the first family, since Y is the normal along this line of intersection. ♦

An exact converse of Dupin's theorem is not true—the lines of intersection of the surfaces of a triple family may be lines of curvature on all the surfaces, even though the surfaces are not orthogonal. For example, the first family may consist of concentric spheres around 0, the second family of planes through the z -axis, and the third family of non-circular cones. On the other hand, if none



of the surfaces in question have umbilics, so that the lines of curvature on each are orthogonal, then the surfaces are clearly orthogonal. The following is a less trivial converse to Theorem 10.

11. THEOREM (DARBOUX). If two families of surfaces are orthogonal, and the intersections are lines of curvature on both, then there exists a third family of surfaces orthogonal to the first two families.

PROOF. Let Δ_1 and Δ_2 be the distributions formed by the tangent planes to the first and second family of surfaces, respectively. Let Δ_3 be the 2-dimension distribution which is everywhere perpendicular to both Δ_1 and Δ_2 . Pick unit vector fields X, Y, Z with

$$X \in \Delta_1 \cap \Delta_3, \quad Y \in \Delta_2 \cap \Delta_3, \quad Z \in \Delta_1 \cap \Delta_3.$$

The hypotheses imply that $\nabla'_X Y = \lambda X$ and $\nabla'_Y X = \mu Y$ for certain functions λ and μ . So

$$[X, Y] = \nabla'_X Y - \nabla'_Y X = \lambda X - \mu Y \in \Delta_3.$$

This shows that Δ_3 is integrable, so the third family of surfaces exists, by the Frobenius integrability theorem. ♦

Dupin's Theorem has as a consequence a geometric proof of a theorem about maps $f: U \rightarrow V$ from an open set $U \subset \mathbb{R}^3$ to an open set $V \subset \mathbb{R}^3$ which are conformal (angle preserving). In the case of maps $f: U \rightarrow V$ with $U, V \subset \mathbb{R}^2 = \mathbb{C}$, it is well-known (Problem 9) that these are precisely the maps which are complex analytic or whose conjugates are. In \mathbb{R}^3 the situation is quite different. One class of conformal maps are the **similarities**, the compositions of translations, orthogonal maps, and multiplications by non-zero constants. There is also an analogue for \mathbb{R}^3 of the complex analytic map $z \mapsto 1/z$ from $\mathbb{C} - \{0\}$ to $\mathbb{C} - \{0\}$. The analogue is easiest to see if we compose this map with conjugation (= reflection through the real axis), so that we obtain the conformal map

$$z \mapsto \frac{1}{\bar{z}} = \frac{z}{|z|^2}.$$

The same formula

$$I(x) = \frac{x}{|x|^2} \quad x \in \mathbb{R}^3 - \{0\},$$

where $|x|$ denotes the norm of x , defines a conformal map (Problem 10), called inversion with respect to the unit sphere. The conformal map

$$x \mapsto r^2 \frac{x}{|x|^2}$$

is called inversion with respect to the sphere of radius r about 0; it keeps points on this sphere fixed, and in general x and $f(x)$ lie on the same line through 0

and $|x| \cdot |f(x)| = r^2$. Of course, we can also consider the inversion

$$x \mapsto x_0 + r^2 \frac{x - x_0}{|x - x_0|^2}$$

with respect to the sphere of radius r about x_0 . Notice that any inversion I' satisfies $I' \circ I' = \text{identity}$ (on its domain).

12. THEOREM (LIOUVILLE). Every conformal map $f: U \rightarrow V$ from a connected open subset U of \mathbb{R}^3 to an open subset V of \mathbb{R}^3 is the restriction to U of a composition of similarities and inversions (in fact, at most one of each).

PROOF. Let $S \subset U$ be any connected surface which is part of a plane or a sphere. We can find a triply orthogonal family of surfaces, with S contained in one of the families, such that the lines of intersection with S are curves with any desired tangent vector at any given point. The image of this triple family under f is again orthogonal, since f is conformal. So by Dupin's Theorem, the lines of intersections of this new family with $f(S)$ are lines of curvature on $f(S)$. Therefore we can find lines of curvature pointing in all directions at any point of $f(S)$. So all points of $f(S)$ are umbilics, and by Theorem 2-2 the surface $f(S)$ is either part of a sphere or part of a plane. We now use

13. LEMMA (MÖBIUS). If $U, V \subset \mathbb{R}^3$ are open sets, with U connected, and $f: U \rightarrow V$ is a map which takes portions of planes and spheres to portions of planes and spheres, then f is the restriction to U of a composition of similarities and immersions (in fact, at most one of each).

PROOF. We begin with a preliminary observation. Let I' be an inversion with respect to a sphere around p , and let S be a sphere with $p \in S$. Then $I'(S - \{p\})$ is a plane. This can be verified by direct calculation, or one can use the following argument: By what we have just shown, $I'(S - \{p\})$ is part of a plane or sphere. It is also easy to see that $I'(S - \{p\})$ is complete, but not compact (for it becomes compact if we add in the point at infinity). So $I'(S - \{p\})$ must be a plane.

Similarly, if P is a plane not containing p , then one can verify by direct calculation that $I'(P)$ is $S - \{p\}$, or one can use the following argument: By what we have just shown, $I'(P)$ is part of a plane or sphere, and contains points arbitrarily close to P . If $I'(P)$ were part of a plane Q , then Q would have to go through p . But I' keeps planes through p fixed, so we would have $P = I'(I'(P)) \subset I'(Q) = Q$, contradicting the fact that P does not contain p .

So $I'(P)$ must be part of a sphere S through p . Since we already know that $I'(S - \{p\})$ is a plane, we easily conclude that $I'(P)$ is all of $S - \{p\}$.

Now to prove the Lemma it obviously suffices to prove that f has the desired form in a neighborhood of any point p , for then f must be analytic, and consequently equal everywhere to any one of these compositions. In particular, we may assume that f is one-one.

Let p_* be a point of U distinct from p , and let Σ_1 be a sphere around p_* such that all points in the ball B bounded by Σ_1 are in U , but $p \notin B$. Let Σ_2 be any sphere around $f(p_*)$. Let $I_1: \mathbb{R}^3 - \{p_*\} \rightarrow \mathbb{R}^3 - \{p_*\}$ be the inversion

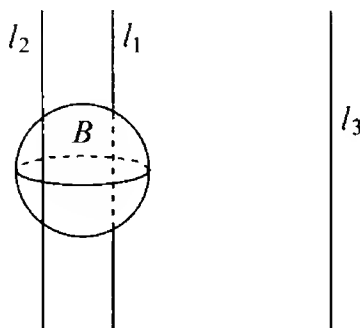


with respect to Σ_1 , and let $I_2: \mathbb{R}^3 - \{f(p_*)\} \rightarrow \mathbb{R}^3 - \{f(p_*)\}$ be the inversion with respect to Σ_2 . Then

$$F = I_2 \circ f \circ I_1: \mathbb{R}^3 - B \rightarrow \mathbb{R}^3$$

is defined everywhere on $\mathbb{R}^3 - B$, has p in its domain, and takes portions of planes and spheres to portions of planes and spheres.

Now if S is any sphere inside Σ_1 with $p_* \in S$, then $f(S)$ must be a sphere in V with $f(p_*) \in f(S)$, so $I_2(f(S) - \{f(p_*)\})$ is a plane. It follows that F takes planes in $\mathbb{R}^3 - B$ into planes of \mathbb{R}^3 . It also follows that F takes straight lines in $\mathbb{R}^3 - B$ into straight lines of \mathbb{R}^3 , since a straight line is the intersection of two planes. We claim that F also preserves parallelism of straight lines. This is clear if l_1 and l_2 are parallel lines lying in a plane $P \subset \mathbb{R}^3 - B$, for then $F(l_1)$ and $F(l_2)$ are disjoint straight lines in $F(P)$. For the case of two parallel lines l_1 and l_2 lying on opposite sides of B , we choose a straight line l_3 parallel

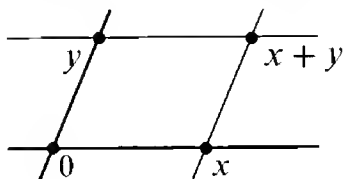


to l_1 and l_2 such that l_1 and l_3 lie in a plane $P_1 \subset \mathbb{R}^3 - B$ while l_2 and l_3 lie in a plane $P_2 \subset \mathbb{R}^3 - B$. Then $F(l_1)$ is parallel to $F(l_3)$ and $F(l_2)$ is parallel to $F(l_3)$, so again $F(l_1)$ is parallel to $F(l_2)$.

Now let T_q denote the translation $x \mapsto x + q$, and consider the map

$$G = T_{-F(p)} \circ F \circ T_p,$$

defined in some convex neighborhood \mathcal{U} of 0. This map takes 0 to 0, and also takes straight lines to straight lines and preserves parallelism. From the parallelogram construction of the sum of two vectors, it is clear that we must



have $G(x + y) = G(x) + G(y)$ whenever x and y are linearly independent vectors with $x, y, x + y \in \mathcal{U}$. The same result holds for linearly dependent x and y , by continuity. From this we easily see that $G(\alpha x) = \alpha G(x)$ for all $\alpha \in \mathbb{R}$ with $x, \alpha x \in \mathcal{U}$. So G is linear, and thus (Problem I.3-31) a composition of an orthogonal map and a self-adjoint map. But G also takes small spheres around 0 to spheres. So we easily see that the self-adjoint factor must be a multiple of the identity, and consequently $G = T_{-F(p)} \circ I_2 \circ f \circ I_1 \circ T_p$ is a similarity in a neighborhood of 0. It follows that f is a composition of similarities and inversions in a neighborhood of p .

To show that f is actually a composition of at most one similarity and inversion, we regard f as extended to the “conformal space” $\mathbb{R}^3 \cup \{\infty\}$, where all similarities are defined at all points, and repeat the proof, choosing $p_* = \infty$. Then the inversion I_1 around p_* is just a similarity on \mathbb{R}^3 , so we obtain a composition of a similarity and one inversion. (If $f(\infty) = \infty$, then I_2 is also a similarity, and our composition reduces to a similarity.)

This completes the proof of the Lemma, and the Theorem. ♦

We already have many results about geodesics, which we obtained in our study of intrinsic Riemannian geometry. The following result, though not at all hard, has always seemed to me particularly nice, because of the way that intrinsic and extrinsic notions are intermingled.

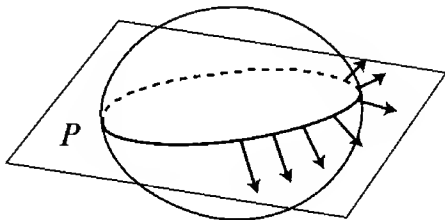
14. THEOREM. Let M be a connected surface in \mathbb{R}^3 such that every geodesic of M is a plane curve. Then M is part of a plane or a sphere.

PROOF. According to Theorem 2-2, it suffices to show that every point $p \in M$ is an umbilic. We can assume that p is not a planar point, since these

are automatically umbilics. Then it certainly suffices to show that any non-asymptotic unit vector $X \in M_p$ is a principal vector. To do this, let c be the geodesic with $c'(0) = X$, lying in the plane P . Proposition 1 shows that the curvature of c is non-zero at 0, and hence in a whole neighborhood of 0. Therefore the desired result follows from

15. LEMMA. If c is a geodesic in M which lies in a plane P and has nowhere 0 curvature, then c is a line of curvature.

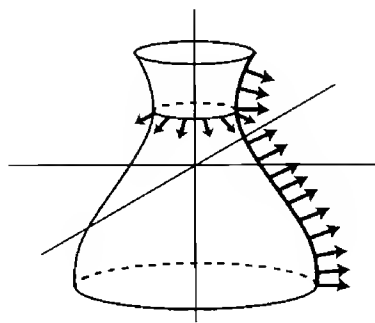
PROOF. Since $c'' \neq 0$ lies in P , and is also perpendicular to the surface, we see that the normal ν to the surface along c lies in P . Hence $d\nu(c(s))/ds$ lies



in P , which means that it must be a multiple of $c'(s)$. [Alternate proof: use Theorem 8.] ♦

To complete our study of geodesics on a surface, we will consider the special case of surfaces of revolution, where the generally intractable differential equations for geodesics reduce to an equation with a simple geometric interpretation. Suppose that our surface is parameterized by

$$f(u, v) = (\rho_1(u) \cos v, \rho_1(u) \sin v, \rho_2(u)),$$



for a curve $\rho = (\rho_1, \rho_2)$ in the (x, z) -plane. Before we do any computations at all, we notice that by Corollary 1-3 the meridians are geodesics, while the parallel at height $\rho_2(u)$ is a geodesic if and only if $\rho_1'(u) = 0$. We will not use all the information given in equations (1) on page 157, but only the fact that the metric has the form

$$\begin{aligned} g_{11}(u, v) &= E(u) \\ g_{12}(u, v) &= 0 \\ g_{22}(u, v) &= G(u) \end{aligned} \quad \implies \quad g^{11} = \frac{1}{E}, \quad g^{12} = 0, \quad g^{22} = \frac{1}{G}.$$

We then compute the Christoffel symbols (as in Chapter 2, the symbol $[ij, k]$ now denotes the Christoffel symbols for the metric $f^*\langle \cdot, \cdot \rangle$ with respect to the usual coordinate system on \mathbb{R}^2):

$$[11, 1] = \frac{1}{2}E'$$

$$[12, 2] = [21, 2] = -[22, 1] = \frac{1}{2}G'$$

$$\text{all other } [ij, k] = 0;$$

$$\Gamma_{11}^1 = \frac{E'}{2E}, \quad \Gamma_{22}^1 = -\frac{G'}{2E}, \quad \Gamma_{12}^2 = \Gamma_{21}^2 = \frac{G'}{2G}$$

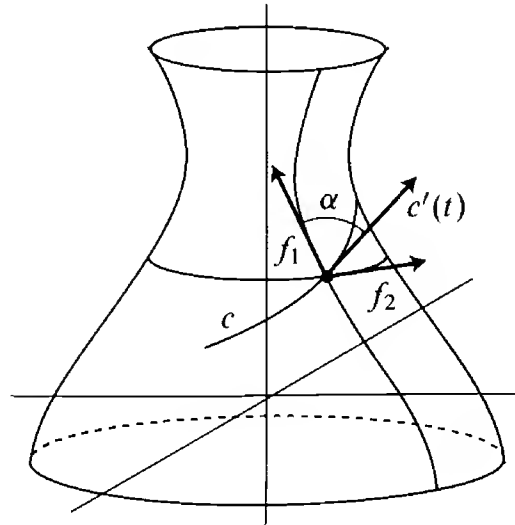
$$\text{all other } \Gamma_{ij}^k = 0.$$

If $t \mapsto \gamma(t) = (\gamma_1(t), \gamma_2(t))$ is a geodesic in \mathbb{R}^2 with the metric $f^*\langle \cdot, \cdot \rangle$, then the equations on pg. I.329 give

$$(*) \quad \frac{d^2\gamma_2}{dt^2} + \frac{G'}{G}(\gamma_1(t)) \frac{d\gamma_1}{dt} \frac{d\gamma_2}{dt} = 0.$$

Now for any curve c on a surface of revolution, let $r(t)$ be the distance from $c(t)$ to the axis of revolution, and let $\alpha(t)$ be the angle between c and the meridian curve that it crosses at time t . More precisely, $\alpha(t)$ is the oriented angle from the meridian tangent vector $f_1(c(t))$ to $c'(t)$ when (f_1, f_2) is chosen as the positive orientation for the tangent space of M , so that

$$\frac{c'}{|c'|} = (\cos \alpha) \frac{f_1}{|f_1|} + (\sin \alpha) \frac{f_2}{|f_2|}.$$



Then equation $(*)$ immediately leads to

16. THEOREM (CLAIRAUT). A geodesic c on a surface of revolution satisfies the equation

$$r(t) \cdot \sin \alpha(t) = A$$

for some constant A . Conversely, if c satisfies this equation and is not a parallel, then c is a geodesic, provided that it is parameterized by arclength.

PROOF. We can write c as $c = f \circ \gamma$ for some curve γ which is a geodesic in \mathbb{R}^2 with the metric $f^*\langle \cdot, \cdot \rangle$. Equation (*) gives

$$\begin{aligned} 0 &= G(\gamma_1(t)) \cdot \gamma_2''(t) + G'(\gamma_1(t)) \cdot \gamma_1'(t) \gamma_2'(t) \\ &= (G \circ \gamma_1)(t) \cdot \gamma_2''(t) + (G \circ \gamma_1)'(t) \cdot \gamma_2'(t) \\ &= [(G \circ \gamma_1) \cdot \gamma_2']'(t), \end{aligned}$$

so $G(\gamma_1(t)) \cdot \gamma_2'(t)$ is constant.

On the other hand, since $g_{12} = 0$, we have

$$\begin{aligned} G(\gamma_1(t)) \cdot \gamma_2'(t) &= \langle f_1(\gamma(t)) \cdot \gamma_1'(t) + f_2(\gamma(t)) \cdot \gamma_2'(t), f_2(\gamma(t)) \rangle \\ &= \langle c'(t), f_2(\gamma(t)) \rangle \\ &= |c'(t)| \cdot |f_2(\gamma(t))| \cdot \sin \alpha(t) \\ &= |c'(t)| \cdot \sqrt{G(\gamma_1(t))} \cdot \sin \alpha(t), \end{aligned}$$

so $\sqrt{G(\gamma_1(t))} \cdot \sin \alpha(t)$ is constant. But the formulas on page 157 show that

$$\sqrt{G(\gamma_1(t))} = \rho_1(\gamma_1(t)),$$

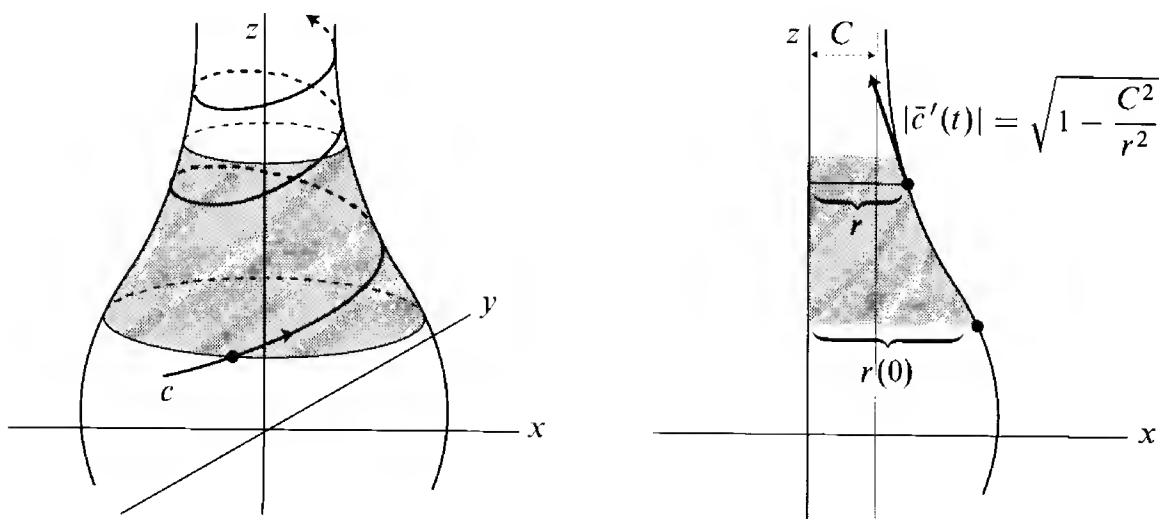
which is exactly the distance of $c(t)$ from the z -axis.

The proof of the converse is left to the reader. ♦

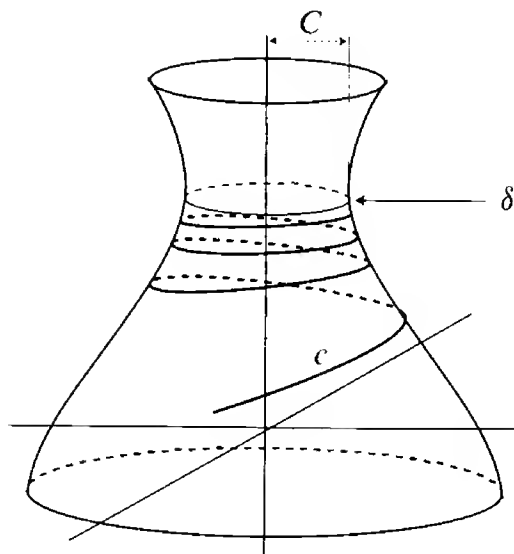
Clairaut's Theorem allows us to give a very complete description of the global behavior of geodesics on surfaces of revolution. Let ρ be the profile curve of the surface, parameterized by arclength, and let \bar{c} denote the projection of c on the (x, z) -plane, so that \bar{c} lies along the image of ρ . Let us suppose that the geodesic c is also parameterized by arclength, and, for concreteness, that $c'(0)$ is pointing upwards. If our geodesic c satisfies $r(t) \cdot \sin \alpha(t) = C$, then the length of $\bar{c}'(t)$ is

$$\begin{aligned} |\bar{c}'(t)| &= \langle \rho'(t), c'(t) \rangle = \cos \alpha(t) \\ &= \sqrt{1 - \frac{C^2}{r(t)^2}}. \end{aligned}$$

From this we see that so long as c lies in a region where $r(t)$ is bounded away from C , the tangent vector $\bar{c}'(t)$ will have length bounded away from 0. It is now easy to deduce the following: If the profile curve ρ never comes within distance C of the z -axis, as we traverse it in the direction of \bar{c} , then \bar{c} must traverse the whole of ρ in this direction.



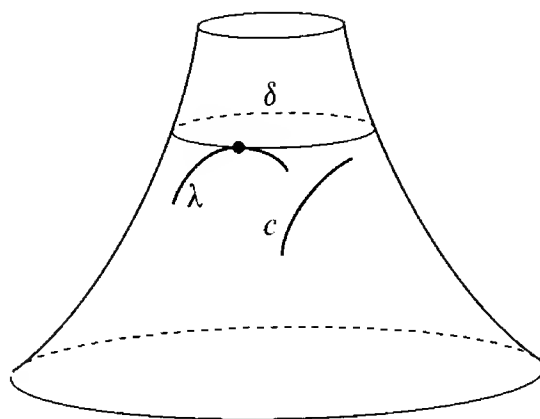
Now suppose that ρ does come within distance C of the z -axis, and let δ be the first meridian above the one at $r(0)$ which has radius C . Then c clearly must come arbitrarily close to δ . If δ happens to be a geodesic, then c cannot intersect δ , for we would then have $\alpha = \pi/2$, which would mean that c' would point along a tangent vector of δ .



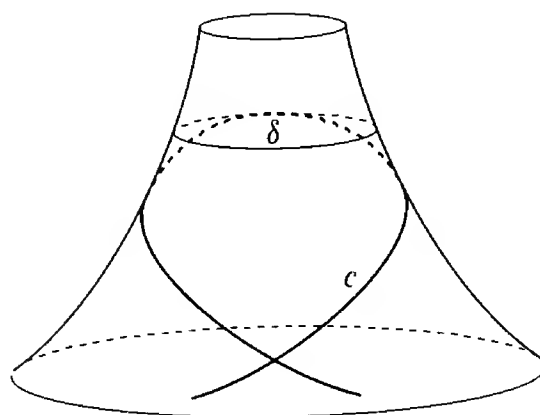
Finally, suppose that δ is *not* a geodesic. Consider a geodesic λ starting at a point of δ , with tangent vector pointing along δ . If β is the angle which λ makes with the meridian, then λ satisfies

$$r(t) \cdot \sin \beta(t) = \text{constant},$$

and $r(t) = C$ when $\beta(t) = \pi/2$, so the constant must also be C . Since δ is not a geodesic, the region directly above δ has $r < C$, and λ cannot go into it. Consequently, it enters the region where c is. A rotation about the axis will



bring λ into coincidence with c since they are both determined by the same constant C , and naturally the rotated curve λ is still a geodesic. In other words, we have shown that c eventually hits δ . Moreover, c' points along δ' at the intersection point (as it must, since $r(t) \cdot \sin \alpha(t) = C$). In addition, c must bounce off δ and proceed downwards. Naturally, the shape of the part going downwards must differ from that of the part going upwards only by a reflection.



ADDENDUM 1

SPECIAL PARAMETER CURVES

Proposition I.5-18 immediately implies

17. COROLLARY. Let p be a point on a surface M in \mathbb{R}^3 .

(a) If p is not an umbilic point, then there is an imbedding $f: U \rightarrow M$, with $p \in f(U)$, whose parameter curves are lines of curvature.

(b) If $K(p) < 0$, then there is an imbedding $f: U \rightarrow M$, with $p \in f(U)$, whose parameter curves are asymptotic curves.

It is sometime useful, especially in the next chapter, to write some of our formulas in terms of these and other special coordinate systems; readers may check for themselves that the following formulas are correct.

A. *The parameter lines are orthogonal.*

Then $F = 0$, and the formula in Problem 13 becomes (subscripts denoting partial derivatives)

$$K = -\frac{1}{2\sqrt{EG}} \left[\left(\frac{E_2}{\sqrt{EG}} \right)_2 + \left(\frac{G_1}{\sqrt{EG}} \right)_1 \right].$$

B. *The parameter lines are lines of curvature.*

In this case, of course, we still have the equation from (A). We also have

$$l = k_1 E, \quad n = k_2 G, \quad m = 0, \quad F = 0.$$

So the Codazzi-Mainardi equations (page 56) become

$$l_2 = \frac{E_2}{2} \left(\frac{l}{E} + \frac{n}{G} \right)$$

$$n_1 = \frac{G_1}{2} \left(\frac{l}{E} + \frac{n}{G} \right).$$

C. *The parameter lines are asymptotic curves.*

We have $l = n = 0$, and the Codazzi-Mainardi equations become

$$m_1 = \frac{\left[\frac{1}{2}(EG - F^2)_1 + FE_2 - EG_1 \right]}{EG - F^2} \cdot m$$

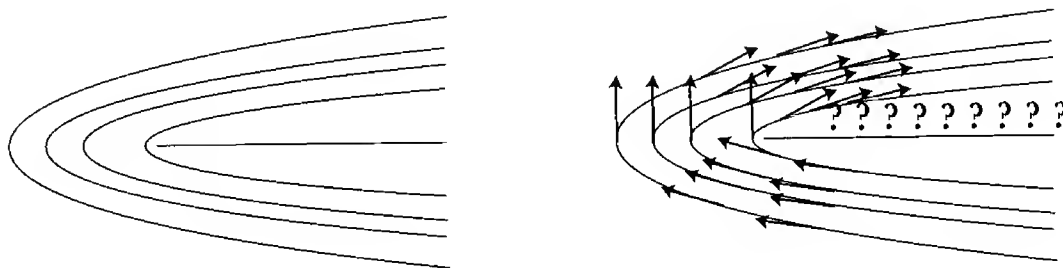
$$m_2 = \frac{\left[\frac{1}{2}(EG - F^2)_2 + FG_1 - GE_2 \right]}{EG - F^2} \cdot m.$$

ADDENDUM 2

SINGULARITIES OF LINE FIELDS

In Chapter I.11 we defined the index of an isolated zero of a vector field, and we proved (Theorem I.11-30) that if a vector field on a compact oriented manifold M has only isolated zeros, then the sum of the indices of these zeros is the Euler characteristic of M .

Now consider the situation where we have a 1-dimensional distribution Δ defined in a neighborhood of a point p of a 2-dimensional manifold M , except at the point p itself. As is easily seen from the pictures below, it may not be



possible to find a nowhere zero vector field X such that $\Delta(q)$ is always spanned by $X(q)$. Nevertheless, we will define an index of Δ at p .

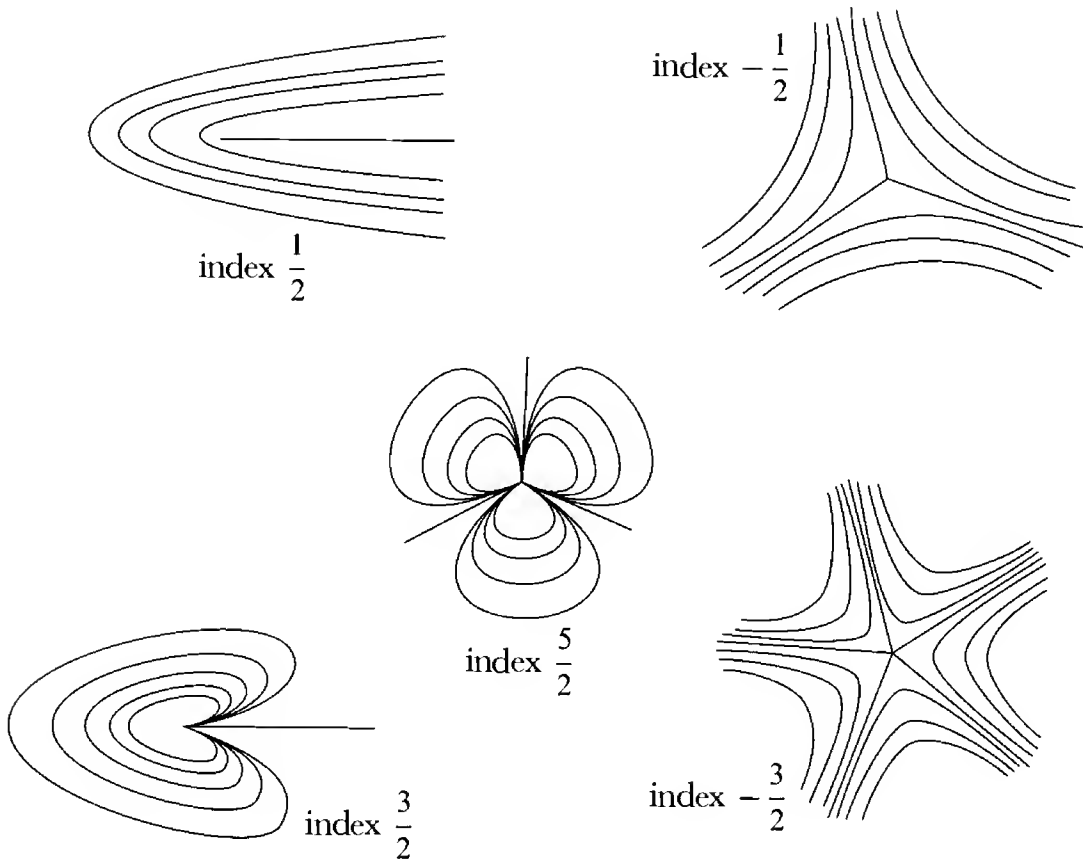
As in the case of a vector field, we first suppose that the distribution Δ is defined on $U - \{0\}$, for U a neighborhood of $0 \in \mathbb{R}^2$. We introduce the projective line \mathbb{P}^1 , which is S^1 with all pairs of antipodal points x and $-x$ identified. Alternatively, \mathbb{P}^1 is the set of all lines through $0 \in \mathbb{R}^2$, and thus the set of all directions in \mathbb{R}^2 . Then we have a map $f_\Delta: U - \{0\} \rightarrow \mathbb{P}^1$ defined by $f_\Delta(q) =$ the direction of $\Delta(q)$. If we let $i: S^1 \rightarrow U$ be $i(x) = \varepsilon x$ for some $\varepsilon > 0$, then we have the map $f_\Delta \circ i: S^1 \rightarrow \mathbb{P}^1$. But \mathbb{P}^1 is homeomorphic to S^1 —we can define a homeomorphism $\alpha: \mathbb{P}^1 \rightarrow S^1$ by noting that \mathbb{P}^1 is the same as a semi-circle with end points identified. Thus we have a map

$$\alpha \circ f_\Delta \circ i: S^1 \rightarrow S^1,$$

and this map has a certain degree. We define the **index** of Δ at p to be

$$\text{index of } \Delta \text{ at } p = \frac{1}{2} \text{degree}(\alpha \circ f_\Delta \circ i).$$

The same arguments which we used in Chapter I.11 allow us to extend this definition from \mathbb{R}^2 to any arbitrary surface. Some examples of these indices are given below.



If Δ happens to be of the form $\Delta(q) = \text{space spanned by } X(q)$, for a vector field X , and $f_X: U - \{0\} \rightarrow S^1$ is the map taking q to $X(q)/|X(q)| \in S^1$, then we have the commutative diagram

$$\begin{array}{ccc}
 S^1 & \xrightarrow{f_X \circ i} & S^1 \\
 & \searrow f_\Delta \circ i & \downarrow \pi \\
 & & \mathbb{P}^1 \\
 & & \downarrow \alpha \\
 & & S^1
 \end{array}$$

where π is the natural projection. Since $\alpha \circ \pi: S^1 \rightarrow S^1$ has degree 2, it follows

that we have

$$\begin{aligned}
 \text{index of } \Delta \text{ at } p &= \frac{1}{2} \text{degree}(\alpha \circ f_{\Delta} \circ i) \\
 &= \frac{1}{2} \text{degree}(\alpha \circ \pi \circ f_X \circ i) \\
 &= \text{degree}(f_X \circ i) \\
 &= \text{index of } X \text{ at } p.
 \end{aligned}$$

In particular, the index of Δ at p is an integer in this case.

Conversely, if the index of Δ at p is an integer, then a suitable vector field X can be found. The easiest way to see this is to give an alternative description of the index. Let $c: [0, 1] \rightarrow \mathbb{R}^2$ be the curve $c(t) = \varepsilon(\cos 2\pi t, \sin 2\pi t)$. The angle between the x -axis and the direction of $\Delta(q)$ is defined only up to a multiple of π [while the angle between the x -axis and a vector is defined only up to a multiple of 2π], but we can find a continuous function $\theta: [0, 1] \rightarrow \mathbb{R}$ such that $\theta(t)$ is an angle between the x -axis and the direction of $\Delta(c(t))$. Then (compare Proposition II.1-6) we have

$$\text{index of } \Delta \text{ at } p = \frac{1}{2\pi}[\theta(1) - \theta(0)].$$

If this index is an integer, so that $\theta(1) - \theta(0)$ is a multiple of 2π , then we can pick out X along c by letting $X(c(t))$ be the unit vector in $\Delta(c(t))$ such that $\theta(t)$ is an angle between $X(c(t))$ and the x -axis.

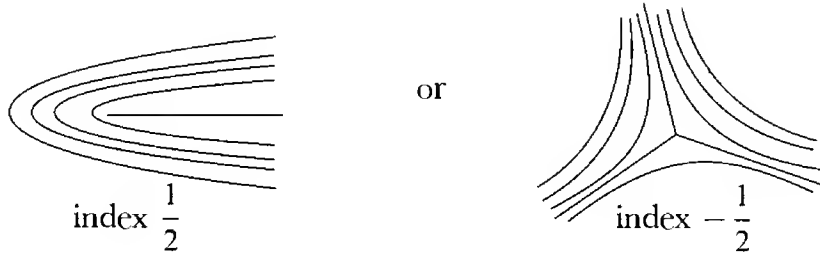
Given any 1-dimensional distribution Δ defined in $M - \{p_1, \dots, p_k\}$, we can define a 2-fold covering space $\varpi: M' \rightarrow M - \{p_1, \dots, p_k\}$ just as on page 198: we let the two points of $\varpi^{-1}(q)$ correspond to the two unit vectors in $\Delta(q)$. If U is an open ball around p_i , then $\varpi^{-1}(U)$ is either two disjoint copies of $U - \{p\}$, or else it is connected and $\varpi|_{\varpi^{-1}(U)}$ looks like the map $z \mapsto z^2$ taking

$$\{z \in \mathbb{C} : 0 < |z| < 1\} \rightarrow \{z \in \mathbb{C} : 0 < |z| < 1\}.$$

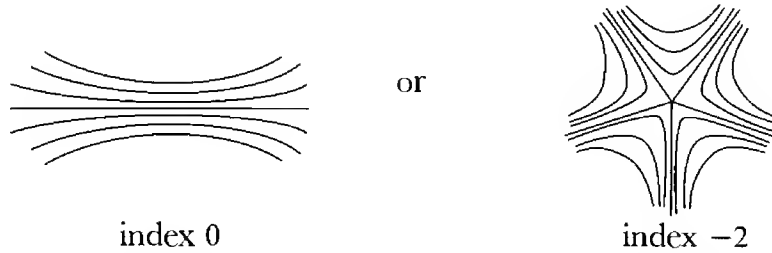
The first case occurs when Δ comes from a vector field, and the second case occurs when it does not. In the former case, we will add two new points to M' , one for each of the disjoint copies of $U - \{p_i\}$, define ϖ of each of these two new points to be p_i , and define the neighborhoods of these new points in the obvious way so that $\varpi^{-1}(U)$ now consists of two disjoint copies of U . In the second case, we will add just *one* new point p_i^* whose neighborhoods we define to be the sets $\{p_i^*\} \cup \varpi^{-1}(A - \{p_i\})$ for A a neighborhood of p_i in U . Let \tilde{M} be M' with all new points added. Then \tilde{M} is a manifold, but the map

$\varpi: \tilde{M} \rightarrow M$ is not a 2-fold covering space, for over certain points p_i it looks like the map $z \mapsto z^2$. These points p_i are called “branch points” of the 2-fold “branched covering space” $\varpi: \tilde{M} \rightarrow M$. There is clearly a distribution $\tilde{\Delta}$ on $\varpi^{-1}(M - \{p_1, \dots, p_k\})$ with $\varpi_*\tilde{\Delta} = \Delta$; moreover $\tilde{\Delta}$ obviously comes from a vector field.

If our original Δ looks like



in a neighborhood U of p_i , then $\tilde{\Delta}$ looks something like



(to see this, just note that if we wrap the bottom pictures twice around the origin, by $z \mapsto z^2$, then the images cover the top pictures). In general,

18. LEMMA. Let $\varpi: \tilde{M} \rightarrow M$ be a 2-fold branched covering space, and let p be a branch point. Let Δ be a 1-dimensional distribution defined on a neighborhood of p , except at p itself, and let $\tilde{\Delta}$ be a 1-dimensional distribution defined on a neighborhood of $\varpi^{-1}(p)$, except at $\varpi^{-1}(p)$ itself, such that $\varpi_*\tilde{\Delta} = \Delta$. Then the index \tilde{i} of $\tilde{\Delta}$ at $\varpi^{-1}(p)$ is related to the index i of Δ at p by

$$\tilde{i} = 2i - 1.$$

PROOF. Regard both p and $\varpi^{-1}(p)$ as $0 \in \mathbb{C}$, and ϖ as the map $z \mapsto z^2$. Let $c: [0, 1] \rightarrow \mathbb{R}^2$ be the semi-circle $c(t) = \varepsilon(\cos \pi t, \sin \pi t)$, and let $\theta: [0, 1] \rightarrow \mathbb{R}$ be a continuous function such that $\theta(t)$ is an angle between the x -axis and the direction of $\tilde{\Delta}(c(t))$. Note that by our construction of $\tilde{\Delta}$ we have

$$\tilde{i} = 2 \cdot \frac{1}{2\pi} [\theta(1) - \theta(0)].$$

Now the curve $\varpi \circ c: [0, 1] \rightarrow \mathbb{R}^2$ is the circle

$$\varpi \circ c(t) = \varepsilon^2(\cos 2\pi t, \sin 2\pi t).$$

It is easy to see that the function

$$\phi(t) = \theta(t) + \pi t$$

gives an angle between the x -axis and the direction of $\Delta(\varpi \circ c(t))$. So

$$\begin{aligned} i &= \frac{1}{2\pi}[\phi(1) - \phi(0)] = \frac{1}{2\pi}[\theta(1) - \theta(0)] + \frac{1}{2} \\ &= \frac{\tilde{i}}{2} + \frac{1}{2}. \quad \blacklozenge \end{aligned}$$

It is also easy to find the Euler characteristic $\chi(\tilde{M})$.

19. LEMMA. If $\varpi: \tilde{M} \rightarrow M$ is a 2-fold branched covering space with l branch points p_1, \dots, p_l , then

$$\chi(\tilde{M}) = 2\chi(M) - l.$$

FIRST PROOF. We will use a triangulation of M (Problem 17 suggests a simple proof that any compact surface can be triangulated). Choose the triangulation so that p_1, \dots, p_l are included among the V vertices (0-simplexes), and let there be E edges (1-simplexes) and F faces (2-simplexes). There is an obvious triangulation of \tilde{M} with 2 vertices over each vertex of M , except for the vertices p_1, \dots, p_l over which there is only 1 vertex. So the number \tilde{V} of vertices of \tilde{M} is

$$\tilde{V} = 2V - l,$$

while we have

$$\tilde{E} = 2E, \quad \tilde{F} = 2F.$$

So

$$\chi(\tilde{M}) = \tilde{V} - \tilde{E} + \tilde{F} = 2V - l - 2E + 2F = 2\chi(M) - l.$$

SECOND PROOF. Let X be a vector field on M with only finitely many zeros q_1, \dots, q_k , and let ι_j be the index of X at q_j . We might as well assume that the p_i are contained among the q 's, for at any point we can always introduce a new zero of X (with index 0). Then there is a vector field \tilde{X} on $\varpi^{-1}(M - \{q_1, \dots, q_k\})$ with $\varpi_*\tilde{X} = X$. If q_j is a branch point, then by Lemma 18,

$$\text{index of } \tilde{X} \text{ at } \varpi^{-1}(q_j) = 2\iota_j - 1.$$

If q_j is not a branch point, then $\varpi^{-1}(q_j)$ consists of two points q'_j, q''_j , and

$$\text{index of } \tilde{X} \text{ at } q'_j \text{ or } q''_j = \iota_j.$$

Then Theorem I.11-30 gives

$$\begin{aligned} \chi(\tilde{M}) &= \text{sum of the indices of } \tilde{X} \\ &= 2 \sum \iota_j - \text{number of branch points} \\ &= 2\chi(M) - l. \quad \spadesuit \end{aligned}$$

Exactly the same sort of reasoning which was used in this second proof leads us to our main result.

20. THEOREM. Let M be a compact oriented surface, and Δ a 1-dimensional distribution on $M - \{p_1, \dots, p_k\}$. Then the sum of the indices of Δ at the p_i is $\chi(M)$.

PROOF. Consider the 2-fold branched covering $\varpi: \tilde{M} \rightarrow M$ constructed previously, and the distribution $\tilde{\Delta}$ on \tilde{M} . If p_i is a branch point, then by Lemma 18

$$\text{index of } \tilde{\Delta} \text{ at } \varpi^{-1}(p_i) = 2(\text{index of } \Delta \text{ at } p_i) - 1.$$

If p_i is not a branch point, then $\varpi^{-1}(p_i)$ consists of two points p'_i, p''_i , and

$$\text{index of } \tilde{\Delta} \text{ at } p'_i \text{ or } p''_i = \text{index of } \Delta \text{ at } p_i.$$

It follows that

$$(1) \quad \begin{aligned} \text{sum of the indices of } \tilde{\Delta} &= 2(\text{sum of the indices of } \Delta) \\ &\quad - \text{number of branch points.} \end{aligned}$$

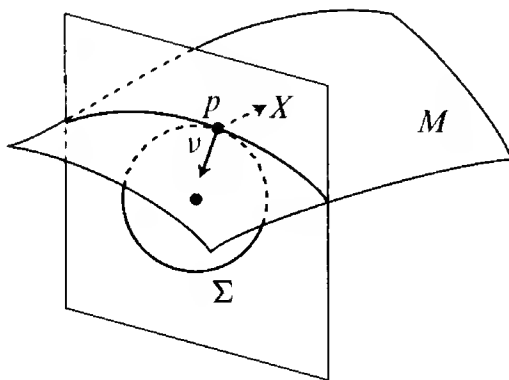
Since $\tilde{\Delta}$ comes from a vector field, it follows from Theorem I.11-30 that

$$(2) \quad \text{sum of the indices of } \tilde{\Delta} = \chi(\tilde{M}).$$

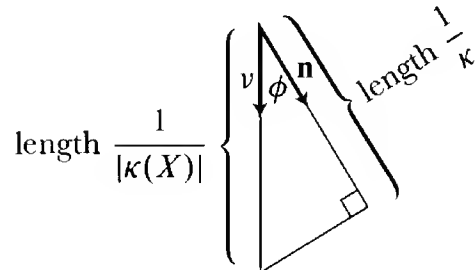
The result now follows from (1), (2), and Lemma 19. \spadesuit

PROBLEMS

1. Let $M \rightarrow \mathbb{R}^3$ be a surface, and $X \in M_p$ a unit vector. Let Σ be the circle in the plane perpendicular to X which is tangent to M at p and has radius $1/|\kappa(X)|$. Show that for every curve c in M with $c'(0) = X$, the center of the



osculating circle of c at 0 lies on Σ . The picture below is a hint.



2. Let $c: \mathbb{R} \rightarrow \mathbb{R}^3$ be a curve which lies on a sphere of radius r .
- (a) We have $\kappa_n = 1/r$, and consequently $\kappa = \sqrt{1/r^2 + \kappa_g^2} > 0$.
- (b) If $\mathbf{t}, \mathbf{n}, \mathbf{b}$ is the Frenet frame of c , and $\mathbf{v}(s) = \nu(c(s))$, where ν is the unit normal of the sphere in which c lies, then

$$\begin{aligned} 0 &= \langle \mathbf{t}, \mathbf{v} \rangle' = \kappa \langle \mathbf{n}, \mathbf{v} \rangle + 1/r \\ \langle \mathbf{n}, \mathbf{v} \rangle' &= \tau \langle \mathbf{b}, \mathbf{v} \rangle \\ \langle \mathbf{b}, \mathbf{v} \rangle' &= -\tau \langle \mathbf{n}, \mathbf{v} \rangle. \end{aligned}$$

- (c) If $\kappa' = 0$, then $\tau = 0$. If κ' is nowhere 0, then τ is nowhere 0 and

$$\frac{\tau}{\kappa} + \left[\frac{1}{\tau} \left(\frac{1}{\kappa} \right)' \right]' = 0.$$

- (d) Conversely, if this condition holds, then c lies on some sphere.
 (e) For convenience, say $r = 1$, so that

$$\kappa = \sqrt{1 + \kappa_g^2}.$$

Let $\mathbf{t}, \mathbf{u}, \mathbf{v}$ be the Darboux frame for c . By differentiating the equations

$$\begin{aligned}\mathbf{t}' &= \kappa_g \mathbf{u} + \mathbf{v} \\ \mathbf{t}' &= \kappa \mathbf{n},\end{aligned}$$

show that

$$\tau = \frac{\kappa_g'}{1 + \kappa_g^2}.$$

3. (a) Let $X \in M_p$ be a unit vector, and suppose that the geodesic γ with $\gamma'(0) = X$ has $\kappa(0) = 0$. Then

$$|\tau_g(X)| = \sqrt{-K(p)}.$$

- (b) If γ is a geodesic with $\gamma'(0) \in M_p$ a principal vector and $\kappa(0) = 0$, then $K(p) = 0$.
 (c) If γ_i are geodesics with perpendicular tangent vectors $\gamma_i'(0) \in M_p$, and $\kappa_1(0) = 0$, but $\kappa_2(0) \neq 0$, then the torsion $\tau_2(0)$ of γ_2 satisfies

$$|\tau_2(0)| = \sqrt{-K(p)}.$$

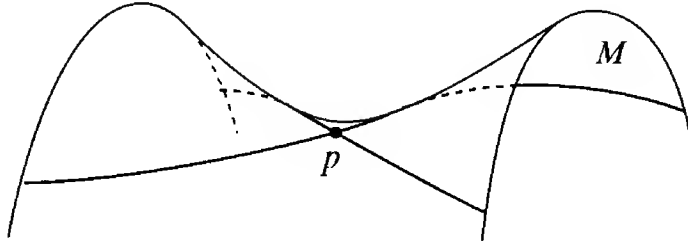
4. (a) Let c be an arclength parameterized curve on a surface $M \subset \mathbb{R}^3$ such that $c'(0) \in M_p$ is an asymptotic direction. If c is *not* asymptotic at p [$c''(0) \notin M_p$], then $\kappa(0) = 0$.

- (b) Now suppose that $c''(0) \in M_p$. Let \bar{c} be the arclength parameterized asymptotic curve with $\bar{c}'(0) = c'(0)$, and denote the curvature and torsion of \bar{c} by $\bar{\kappa}$ and $\bar{\tau}$.

Using Laguerre's formulation of Proposition 4, and equation (13) on page 192, show that

$$\begin{aligned}\kappa(0)[3\bar{\tau}(0) - \tau(0)] &= 2\bar{\tau}(0)\bar{\kappa}(0) \\ \implies \kappa(0)[\pm 3\sqrt{-K(p)} - \tau(0)] &= \pm 2\sqrt{-K(p)}\bar{\kappa}(0).\end{aligned}$$

(c) Show that at a point $p \in M$ where $K(p) < 0$, the intersection of M and the tangent plane at p consists of two curves which cross each other at p , and which point in the asymptotic directions. For either of these curves, show that



the curvature at p is $2/3$ times the curvature of the corresponding asymptotic curve through p (Beltrami).

5. Let $M \subset \mathbb{R}^3$ be a surface, and let X_1, X_2 be an orthonormal moving frame, in terms of which we write

$$\Pi = \sum_{i,j} l_{ij} \theta^i \otimes \theta^j.$$

(a) By Problem 2-5 we have

$$\nabla \Pi = \sum_{i,j,k} l_{ij;k} \theta^i \otimes \theta^j \otimes \theta^k$$

where

$$\sum_k l_{ij;k} \theta^k = dl_{ij} - \sum_\rho l_{\rho j} \omega_i^\rho - \sum_\rho l_{i\rho} \omega_j^\rho.$$

If X_1, X_2 is the frame X, \bar{X} in the proof of Proposition 4, apply this equation to X_1 to obtain

$$(1) \quad l_{11;1} = \frac{dl_{11}}{ds} - 2l_{12} \cdot \omega_1^2(X_1),$$

and deduce the first part of Proposition 4. Deduce the second part similarly.

(b) Show that

$$d\psi_i^3 = \sum_j dl_{ij} \wedge \theta^j - \sum_{j,\rho} l_{ij} \omega_\rho^j \wedge \theta^\rho.$$

Conclude from equation (1) that

$$(l_{12;1} - l_{11;2})\theta^1 \wedge \theta^2 = 0.$$

(c) If

$$\nabla \nabla \Pi = \sum_{i,j,k,h} l_{ij;kh} \theta^i \otimes \theta^j \otimes \theta^k \otimes \theta^h,$$

show that

$$\begin{aligned} l_{11;11} &= \frac{dl_{11;1}}{ds} - 3l_{12;1}\omega_1^2(X_1) \\ l_{12;11} &= \frac{dl_{12;1}}{ds} - (2l_{22;1} - l_{11;1})\omega_1^2(X_1) \\ &= \frac{dl_{12;1}}{ds} - \left(2\left[2\frac{dH}{ds} - l_{11;1}\right] - l_{11;1}\right)\omega_1^2(X_1). \end{aligned}$$

(d) For all arclength parameterized curves c in M with the same tangent vector $c'(0) \in M_p$ the quantity

$$\kappa_n''(s) - 2\tau_g(s)\kappa_g'(s) - 5\tau_g'(s)\kappa_g(s) - 6[\kappa_n(s) - H(c(s))]\kappa_g^2(s)$$

has the same value at $s = 0$. The same is true for

$$\begin{aligned} \tau_n''(s) + [2\kappa_n(s) - H(c(s))]\kappa_g'(s) \\ + \left[5\kappa_n'(s) - 6\tau_g(s)\kappa_g(s) - 6\frac{dH(c(s))}{ds}\right]\kappa_g(s). \end{aligned}$$

6. Let $X \in M_p$ be a principal vector, and c the principal curve with $c'(0) = X$.

- (a) Determine $\kappa_n'(0)$.
- (b) Use Problem 5 to determine $\kappa_g'(0)$ [in terms of $X(H)$].
- (c) Show how to determine $\kappa'(0)$.
- (d) Show how to use equation (7) on page 189 to determine $\phi'(0)$, and then how to find $\tau(0)$.

7. Let $\pi: \tilde{M} \rightarrow M$ be an n -fold covering of a compact orientable manifold M .

- (a) Let X be a vector field on M with only finitely many zeros. Show that there is a vector field \tilde{X} on \tilde{M} with n zeros of index ι for every zero of X with index ι . Conclude that $\chi(\tilde{M}) = n \cdot \chi(M)$.
- (b) Also prove this result by finding a triangulation of M for which there is a corresponding triangulation of \tilde{M} with n k -simplexes for each k -simplex of M .

8. Equation (*) on page 205 is the basis for the best coordinate system for ellipsoids and hyperboloids of one or two sheets. For given (x, y, z) let $\phi(\lambda)$ be the cubic

$$(1) \quad \phi(\lambda) = (a^2 - \lambda)(b^2 - \lambda)(c^2 - \lambda)(g(\lambda) - 1),$$

with roots

$$\lambda_1 < a^2, \quad a^2 < \lambda_2 < b^2, \quad b^2 < \lambda_3 < c^2.$$

Clearly

$$\phi(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)(\lambda_3 - \lambda).$$

(a) Substituting for the expression for $\phi(\lambda)$ from (1), and choosing $\lambda = a^2, b^2$, and c^2 , obtain

$$x^2 = \frac{(a^2 - \lambda_1)(a^2 - \lambda_2)(a^2 - \lambda_3)}{(a^2 - b^2)(a^2 - c^2)}$$

$$y^2 = \frac{(b^2 - \lambda_1)(b^2 - \lambda_2)(b^2 - \lambda_3)}{(b^2 - a^2)(b^2 - c^2)}$$

$$z^2 = \frac{(c^2 - \lambda_1)(c^2 - \lambda_2)(c^2 - \lambda_3)}{(c^2 - a^2)(c^2 - b^2)}.$$

(b) Setting one $\lambda_i = \text{constant}$, these equations give a parameterization of the surface $g(\lambda_i) = 1$ by means of the two other variables λ_j, λ_k . Setting

$$\begin{aligned} a^2 - \lambda_i &= \alpha, & b^2 - \lambda_i &= \beta, & c^2 - \lambda_i &= \gamma \\ \lambda_j - \lambda_i &= u, & \lambda_k - \lambda_i &= v, \end{aligned}$$

we have the surface

$$\frac{x^2}{\alpha} + \frac{y^2}{\beta} + \frac{z^2}{\gamma} = 1$$

parameterized by

$$x = \sqrt{\frac{\alpha(\alpha - u)(\alpha - v)}{(\alpha - \beta)(\alpha - \gamma)}}$$

$$y = \sqrt{\frac{\beta(\beta - u)(\beta - v)}{(\beta - \alpha)(\beta - \gamma)}}$$

$$z = \sqrt{\frac{\gamma(\gamma - u)(\gamma - v)}{(\gamma - \alpha)(\gamma - \beta)}}.$$

Note that the u - and v -parameter lines are clearly lines of curvature. Calculate that

$$E = \frac{u(u-v)}{f(u)} \quad F = 0 \quad G = \frac{v(v-u)}{f(v)}$$

where $f(t) = 4(\alpha - t)(\beta - t)(\gamma - t)$.

9. Let $f: \mathbb{C} \rightarrow \mathbb{C}$ be

$$f(x, y) = f(x + iy) = (u(x, y), v(x, y)) = u(x, y) + iv(x, y)..$$

(a) Calculate that

$$\begin{aligned} f^*(dx \otimes dx + dy \otimes dy) &= (u_x^2 + v_x^2) dx \otimes dx + (u_y^2 + v_y^2) dy \otimes dy \\ &\quad + (u_x u_y + v_x v_y)(dx \otimes dy + dy \otimes dx). \end{aligned}$$

Conclude that f is conformal if and only if

$$u_x^2 + v_x^2 = u_y^2 + v_y^2 \quad \text{and} \quad u_x u_y + v_x v_y = 0.$$

(b) Show that f is conformal if and only if

$$u_x = \pm v_y, \quad u_y = \mp v_x.$$

Hint: Multiply the first equation of part (a) by v_y^2 .

10. Consider the map

$$I(x) = \frac{x}{|x|^2} \quad x \in \mathbb{R}^3 - \{0\}.$$

(a) If $S^2(r)$ is the 2-sphere around 0 of radius r , and $X \in S^2(r)_p$, then $I_*(X) \in S^2(\frac{1}{r})_{I(p)}$ is parallel to X , and $|I_*(X)| = \frac{1}{r}|X|$.

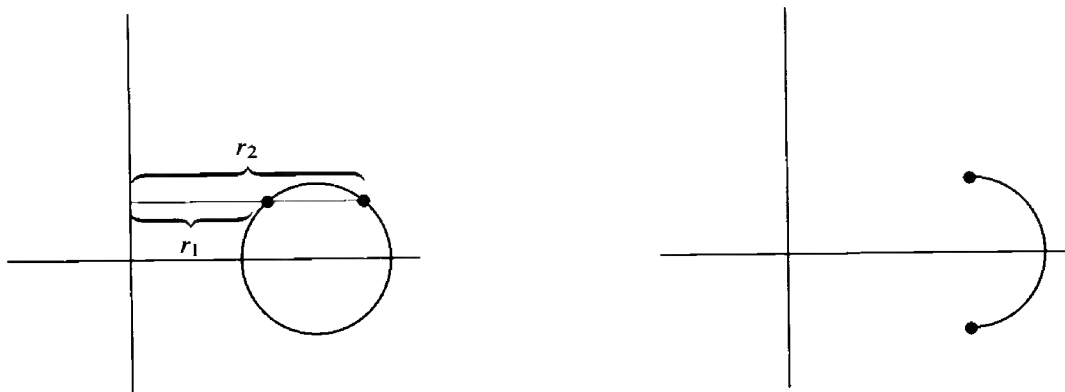
(b) If v is the unit normal to $S^2(r)$ at p , then $I_*(v)$ is $\frac{1}{r}$ times the unit normal to $S^2(\frac{1}{r})$ at $I(p)$.

(c) I is conformal.

11. Note that Lemma 13 is valid also for $U, V \subset \mathbb{R}^2$, where $f: U \rightarrow V$ takes portions of straight lines and circles to portions of straight lines and circles. Conclude that if f is orientation preserving, then f is of the form

$$f(z) = \frac{az + b}{cz + d}, \quad a, b, c, d \in \mathbb{C}.$$

12. Use Clairaut's Theorem to analyze the geodesics on a torus. (The fact that points with different values of r can have the same height is confusing, but irrelevant; it may help to think of the torus in terms of the profile curve on the right, but with the end points identified.)



Answer: All geodesics, except the inner parallel, intersect the outer parallel. Those which intersect at a small angle θ describe a sine-like curve between two parallels at equal distances from the outer one. For a certain angle θ we obtain a sine-like curve between the top and bottom parallels. For somewhat larger θ the curve flops over the top and bottom and bounces between two parallels on the inner part of the torus. For a certain angle θ we obtain a curve which approaches the inner parallel asymptotically. For slightly larger θ the geodesic hits the inner parallel, at a small angle, after going around many times. As θ increases, the geodesic hits the inner parallel at larger angles, after going around fewer times, until, at $\theta = \pi/2$, we obtain a meridian circle.

13. From pg. II.131 we have the formula

$$W^4 K = \det \begin{pmatrix} -\frac{1}{2}G_{11} + F_{12} - \frac{1}{2}E_{22} & \frac{1}{2}E_1 & F_1 - \frac{1}{2}E_2 \\ F_2 - \frac{1}{2}G_1 & E & F \\ \frac{1}{2}G_2 & F & G \end{pmatrix} \\ - \det \begin{pmatrix} 0 & \frac{1}{2}E_2 & \frac{1}{2}G_1 \\ \frac{1}{2}E_2 & E & F \\ \frac{1}{2}G_1 & F & G \end{pmatrix},$$

where $W = \sqrt{EG - F^2}$. Verify the following (seemingly non-rational) expression for K , due to Frobenius:

$$K = -\frac{1}{4W^4} \det \begin{pmatrix} E & E_1 & E_2 \\ F & F_1 & F_2 \\ G & G_1 & G_2 \end{pmatrix} + \frac{1}{2W} \left[\left(\frac{F_2 - G_1}{W} \right)_1 + \left(\frac{F_1 - E_2}{W} \right)_2 \right]$$

14. Let $f: U \rightarrow M$ be an imbedding whose parameter curves are asymptotic curves, so that

$$f_{11} = Af_1 + Bf_2, \quad f_{22} = Cf_1 + Df_2.$$

Choose the orientation so that $\det(f_1, f_2, f_{12}) > 0$.

(a) For the affine first fundamental form \mathfrak{I}_f we have

$$g_{11} = g_{22} = 0 \quad \text{and} \quad g_{12} = \sqrt{\det(f_1, f_2, f_{12})} = \alpha, \text{ say.}$$

(b) The Christoffel symbols for \mathfrak{I}_f are

$$[11, 2] = \alpha_1, \quad [22, 1] = \alpha_2, \quad \text{all other } [ij, k] = 0.$$

$$\Gamma_{11}^1 = \frac{\alpha_1}{\alpha}, \quad \Gamma_{22}^2 = \frac{\alpha_2}{\alpha}, \quad \text{all other } \Gamma_{ij}^k = 0,$$

so

$$\nabla_{f_1} f_1 = \frac{\alpha_1}{\alpha} f_1, \quad \nabla_{f_2} f_2 = \frac{\alpha_2}{\alpha} f_2.$$

(c) We have

$$\mathfrak{A}(f_i, f_j) = f_{ij} - \nabla_{f_i} f_j \implies \ell_{ijk} = \langle f_{ij} - \nabla_{f_i} f_j, f_k \rangle,$$

and consequently

$$\ell_{111} = B\alpha, \quad \ell_{112} = A\alpha - \alpha_1, \quad \ell_{221} = D\alpha - \alpha_2, \quad \ell_{222} = C\alpha.$$

(d) Show that

$$\begin{aligned} (\ell_{111})^2 &= \det(f_1, f_{11}, f_{111}) \\ -(\ell_{222})^2 &= \det(f_2, f_{22}, f_{222}). \end{aligned}$$

[This gives another proof to the second part of Theorem 7.]

(e) Show that

$$\ell_{112} = \ell_{221} = 0.$$

Thus

$$\mathfrak{I}_f = \sqrt{\det(f_1, f_{11}, f_{111})} ds^1 \otimes ds^1 \otimes ds^1 + \sqrt{-\det(f_2, f_{22}, f_{222})} ds^2 \otimes ds^2 \otimes ds^2.$$

15. (a) For the parameterization in Problem 14, show that the Pick invariant is

$$J = \frac{\ell_{111}\ell_{222}}{(g_{12})^3}.$$

(b) On a region of M where $\ell_{111} = 0$, we have $f_{11} = af_1$. Hence M is a ruled surface. Similarly on a region where $\ell_{222} = 0$. Conversely, a ruled surface has $J = 0$.

16. As an alternative to the approach taken in Problem 3-11, use Problem 14 to show that a doubly ruled surface with $K < 0$ has $\mathfrak{T} = 0$, so that it is quadratic, by Proposition 2-19.
17. Show that a compact Riemannian 2-manifold can be triangulated by choosing the vertices to be an ε -dense set, where every point has a geodesically convex neighborhood of radius $> \varepsilon$, and choosing the edges to be geodesic segments.

CHAPTER 5

COMPLETE SURFACES OF CONSTANT CURVATURE

We have already seen, in Chapter 3, that there are many surfaces with constant curvature. On the other hand, few of our examples were complete manifolds. In this chapter we will determine precisely which surfaces in \mathbb{R}^3 can be obtained by isometrically immersing complete manifolds with constant curvature $K > 0$, $K = 0$, or $K < 0$.

In the case of complete surfaces of constant curvature $K > 0$ we will actually assume that the surface is compact; in Chapter 8 (Theorem 8-17), however, we will see that this additional hypothesis is superfluous. By Hadamard's Theorem, our surface is an imbedded submanifold $M \subset \mathbb{R}^3$.

1. LEMMA (HILBERT). Let M be a surface immersed in \mathbb{R}^3 , and let $p \in M$ be a non-umbilic point. Let $k_1 \geq k_2$ be the two principal curvatures on M and suppose that k_1 has a local maximum at p , and k_2 has a local minimum at p . Then $K(p) \leq 0$.

PROOF. According to Addendum 1 to Chapter 4, we can choose an imbedding $f: U \rightarrow M$, with $p \in f(U)$, whose coordinate lines are the lines of curvature. Then Gauss' equation and the Codazzi-Mainardi equations become [subscripts, except those on k_1 and k_2 , denote partial derivatives]

$$(1) \quad K = -\frac{1}{2\sqrt{EG}} \left[\left(\frac{E_2}{\sqrt{EG}} \right)_2 + \left(\frac{G_1}{\sqrt{EG}} \right)_1 \right]$$

$$(2) \quad l_2 = \frac{E_2}{2} \left(\frac{l}{E} + \frac{n}{G} \right) = \frac{E_2}{2} (k_1 + k_2)$$

$$(3) \quad n_1 = \frac{G_1}{2} \left(\frac{l}{E} + \frac{n}{G} \right) = \frac{G_1}{2} (k_1 + k_2);$$

the second equalities in (2) and (3) follow from the fact that

$$l = k_1 E, \quad n = k_2 G.$$

Moreover, differentiation of these last two equations yields

$$l_2 = \frac{\partial k_1}{\partial t} E + k_1 E_2, \quad n_1 = \frac{\partial k_2}{\partial s} G + k_2 G_1.$$

(The functions k_i are differentiable near p , since the functions H and K are differentiable, and $k_i = H \pm \sqrt{H^2 - K}$, where $H^2 - K > 0$ in a neighborhood of the non-umbilic point p .) Together with (2) and (3) we then have

$$(2') \quad E_2 = -\frac{2E}{k_1 - k_2} \cdot \frac{\partial k_1}{\partial t}$$

$$(3') \quad G_1 = \frac{2G}{k_1 - k_2} \cdot \frac{\partial k_2}{\partial s}.$$

Substituting (2'), (3') into (1) gives

$$(1') \quad K = -\frac{1}{2EG} \left[-\frac{2E}{k_1 - k_2} \cdot \frac{\partial^2 k_1}{\partial t^2} + \frac{2G}{k_1 - k_2} \cdot \frac{\partial^2 k_2}{\partial s^2} \right] \\ + (\text{something continuous}) \cdot \frac{\partial k_1}{\partial t} \\ + (\text{something continuous}) \cdot \frac{\partial k_2}{\partial s}.$$

Since k_1 has a local maximum at p , and k_2 a local minimum, we have

$$\frac{\partial k_1}{\partial t}(p) = \frac{\partial k_2}{\partial s}(p) = 0, \quad \frac{\partial^2 k_1}{\partial t^2}(p) \leq 0, \quad \frac{\partial^2 k_2}{\partial s^2}(p) \geq 0.$$

Together with (1') this shows that $K(p) \leq 0$. ♦

2. THEOREM. If M is a compact connected surface in \mathbb{R}^3 with constant curvature $K > 0$, then M is a sphere.

PROOF. Let $k_1 \geq k_2$ be the principal curvatures on M , and let p be a point where k_1 achieves its maximum. Then $k_2 = K/k_1$ has its minimum at p . If we had $k_1(p) > k_2(p)$, so that p was not an umbilic, then the Lemma would imply that $K(p) \leq 0$, a contradiction. Hence $k_1(p) = k_2(p)$. Moreover, for any point $q \in M$ we then have

$$k_1(p) \geq k_1(q) \geq k_2(q) \geq k_2(p) = k_1(p),$$

so also $k_1(q) = k_2(q)$. Thus all points of M are umbilics, and Theorem 2-2 applies. ♦

We also have another result of interest:

3. **THEOREM.** If M is a compact connected surface in \mathbb{R}^3 , with K everywhere > 0 , and constant mean curvature H , then M is a sphere.

PROOF. As before, if k_1 achieves its maximum at p , then $k_2 = H - k_1$ achieves its minimum. Since we are assuming $K(p) > 0$, we find that $k_1(p) = k_2(p)$, and the rest of the proof proceeds as before. ♦

We now turn our attention to immersed surfaces M in \mathbb{R}^3 which are flat, that is, which have $K = 0$ everywhere. We know that if a connected open set $U \subset M$ consists entirely of planar points, then U is part of a plane. Let us consider a point of M which is *not* a planar point, and hence a parabolic point. In a neighborhood U of this point we can choose a C^∞ unit vector field X_1 such that each $X_1(q)$ is a principal vector with principal curvature $k_1(q) = 0$, and another C^∞ unit vector field X_2 , orthogonal to X_1 , such that each $X_2(q)$ is a principal vector with principal curvature $k_2(q) \neq 0$.

4. **PROPOSITION.** The integral curves of X_1 are straight line segments. Consequently, every non-planar point in a flat surface has a neighborhood which is a ruled surface.

PROOF. Let $\theta^i, \omega_1^2, \psi_i^3$ be the forms associated with the adapted orthonormal moving frame (X_1, X_2, ν) . Since

$$\nabla'_{X_i} \nu = \begin{cases} 0 & i = 1 \\ k_2 X_2 & i = 2, \end{cases}$$

we have

$$\begin{aligned} \psi_1^3(X_i) &= -\psi_3^1(X_i) = -\langle X_1, \nabla'_{X_i} \nu \rangle = 0 \\ \psi_2^3(X_i) &= -\psi_3^2(X_i) = -\langle X_2, \nabla'_{X_i} \nu \rangle = -k_2 \theta^2(X_i). \end{aligned}$$

In short, we have

$$\psi_1^3 = 0, \quad \psi_2^3 = -k_2 \theta^2,$$

so one of the Codazzi-Mainardi equations (page 70) gives

$$0 = d\psi_1^3 = -\psi_2^3 \wedge \omega_1^2 = k_2 \theta^2 \wedge \omega_1^2.$$

Since k_2 is never 0, this can happen only if ω_1^2 is always a multiple of θ^2 . This implies that

$$0 = \omega_1^2(X_1) = \langle \nabla'_{X_1} X_1, X_2 \rangle,$$

while we also have

$$0 = \psi_1^3(X_1) = \langle \nabla'_{X_1} X_1, X_3 \rangle.$$

Therefore $\nabla'_{X_1} X_1 = 0$, which means that the integral curves of X_1 are straight lines in \mathbb{R}^3 . ♦

We still haven't said what happens at a planar point which is a limit of parabolic points, but before worrying about such points, we will first obtain some information about flat ruled surfaces, classically known as **developable surfaces**.

Consider a ruled surface

$$(1) \quad f(s, t) = c(s) + t\delta(s),$$

where we assume $|\delta| = 1$ for convenience, but do not necessarily insist on the canonical parameterization (since it is not always possible to introduce it). As we have seen (page 147 and page 197), this surface is flat precisely when c', δ, δ' are everywhere linearly dependent. Let us first consider an open interval for s on which δ, δ' alone are *everywhere linearly dependent*. Since $|\delta| = 1$, we have $\langle \delta, \delta' \rangle = 0$, so actually δ' must be 0, and δ is constant. We then have a portion of a cylinder. Next let us consider an interval on which δ, δ' are *everywhere linearly independent*. Then there are unique C^∞ functions α, β with

$$(2) \quad c'(s) = \alpha(s)\delta(s) + \beta(s)\delta'(s).$$

Let

$$(3) \quad c^*(s) = c(s) - \beta(s)\delta(s).$$

Then

$$(4) \quad \begin{aligned} c^{*'}(s) &= c'(s) - \beta(s)\delta'(s) - \beta'(s)\delta(s) \\ &= [\alpha(s) - \beta'(s)]\delta(s). \end{aligned}$$

Again, we will consider only two special cases. On an interval where $\alpha(s) - \beta'(s)$ is *always* 0, we have $c^*(s) = \text{constant vector } c_0^*$, and by (1) and (3) our surface is

$$f(s, t) = c_0^* + (t + \beta(s))\delta(s),$$

which is a portion of the cone

$$g(s, t) = c_0^* + t\delta(s).$$

On the other hand, on an interval where $\alpha(s) - \beta'(s)$ is *never* 0, we have

$$(5) \quad \delta(s) = \frac{c^{*'}(s)}{\alpha(s) - \beta'(s)},$$

so by (1) and (3) our surface is

$$\begin{aligned} f(s, t) &= c(s) + t\delta(s) = c^*(s) + (t + \beta(s))\delta(s) \\ &= c^*(s) + \left[\frac{t + \beta(s)}{\alpha(s) - \beta'(s)} \right] c^{*'}(s), \end{aligned}$$

which is a portion of the tangent developable

$$g(s, t) = c^*(s) + tc^{*'}(s)$$

of the curve c^* . [Notice that by (4) we have

$$c^{*''}(s) = (\alpha(s) - \beta'(s))\delta'(s) + (\alpha'(s) - \beta''(s))\delta(s);$$

since we are on an interval where δ' and δ are linearly independent and $\alpha - \beta'$ is nowhere 0, this shows that the curve c^* does indeed have non-vanishing curvature on the interval.]

The discussion in the preceding paragraph constitutes the classical “classification” of developable surfaces, which was commonly expressed by saying that all developables are planes, cylinders, cones, or tangent developables. We have clearly not proved any such result, since we have only considered special intervals on which certain conditions hold. Nowadays people tend to say: Oh well, the classical classification of developables was really for analytic surfaces—one ought to say that a connected *analytic* developable surface is either a plane, cylinder, cone, or tangent developable. But *even this* is not true. It is true that if a connected analytic developable surface contains a plane, cylinder or cone, then it must be a plane, cylinder, or cone; for planes, cylinders, and cones are the surfaces which arise in our analysis when certain functions are zero on a whole interval. But an analytic developable surface can also be made up of several tangent developables joined together along a line belonging to neither. For example, consider the analytic function $\delta: \mathbb{R} \rightarrow S^2$ defined by

$$(1) \quad \delta(s) = \frac{1}{\sqrt{(1+s^2)^2 + (1+s^3)^2 + s^8}} \cdot (1+s^2, 1+s^3, s^4).$$

We clearly have

$$(2) \quad \delta'(s) = sA(s)$$

for some analytic function A , and we easily check that

$$(3) \quad A(0) \text{ is not a multiple of } \delta(0).$$

Now let $c: \mathbb{R} \rightarrow \mathbb{R}^3$ be an analytic curve with $c(0) = 0$ and

$$(4) \quad c'(s) = \delta(s) + A(s).$$

Since $c'(0) = \delta(0) + A(0)$ is linearly independent of $\delta(0)$ [by (3)], the map

$$f(s, t) = c(s) + t\delta(s)$$

is an immersion at $(0, t)$ for all t . We claim that f is flat at all points, i.e., that $c'(s), \delta(s), \delta'(s)$ are linearly dependent for all s . This is clear for $s = 0$, since $\delta'(0) = 0$, while for $s \neq 0$ we have, by (4),

$$c'(s) = 1 \cdot \delta(s) + \frac{1}{s}\delta'(s).$$

This equation, together with equation (3) on page 236, shows that for $s \neq 0$ our surface is the tangent developable of the curve

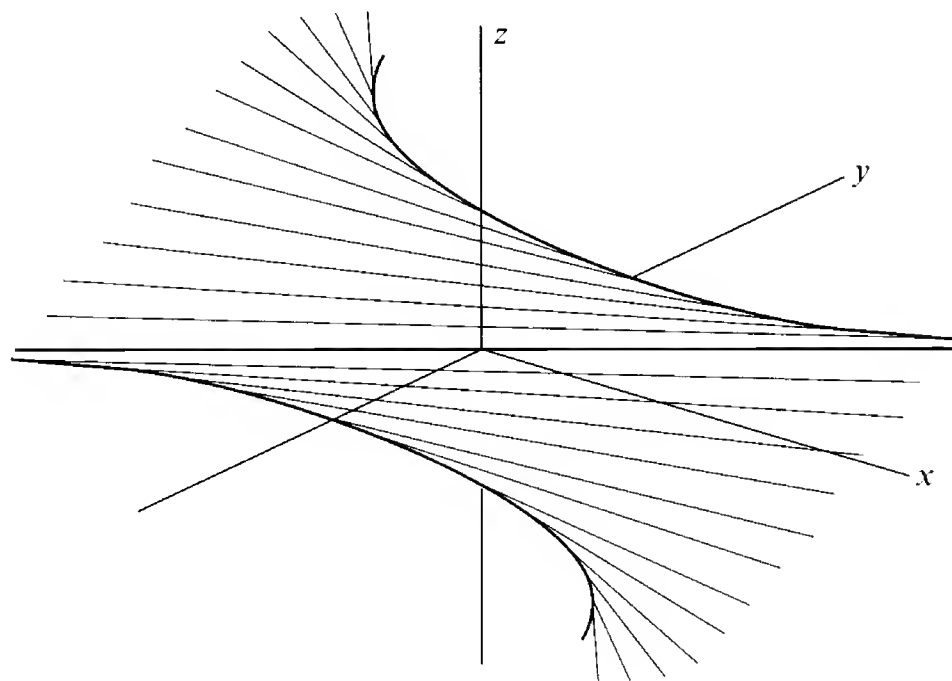
$$c^*(s) = c(s) - \frac{1}{s}\delta(s).$$

Since c and δ are analytic, c^* is definitely *not* analytic at 0. Instead we have two different curves for $s > 0$ and $s < 0$; it is easy to see that

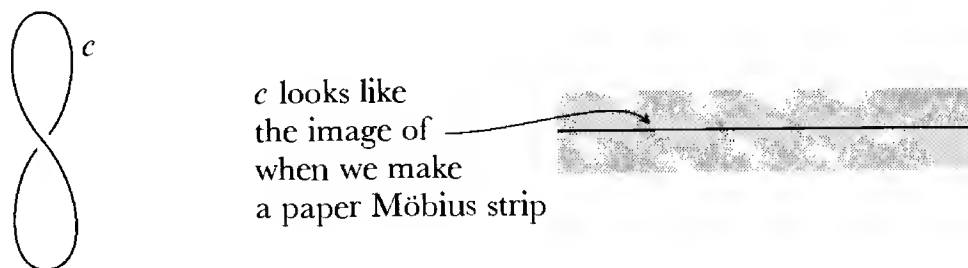
$$c^*(s) \rightarrow (-\infty, -\infty, 0) \quad \text{as } s \rightarrow 0^+$$

$$c^*(s) \rightarrow (\infty, \infty, 0) \quad \text{as } s \rightarrow 0^-.$$

Our surface looks something like the following picture:

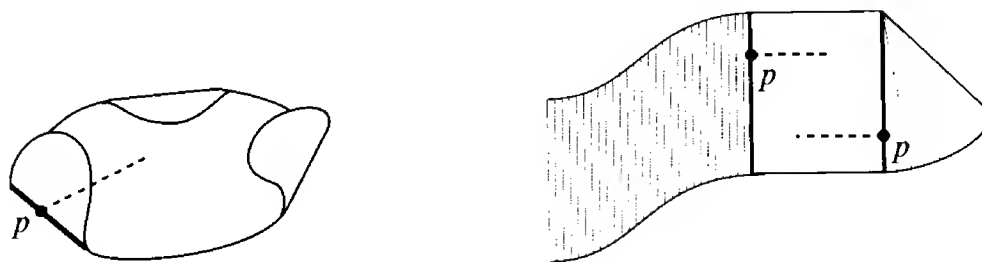


The possible complexity of analytic developables is perhaps most strikingly illustrated by the fact that there is an *analytic* developable surface which is homeomorphic to the Möbius strip; it may be constructed as follows (see Wunderlich [1] for details). We take an analytic closed curve c which looks like the center line of the paper Möbius strip. In the paper model, this line is a geodesic,



desic, since it comes from a straight line in the flat piece of paper which we bent to form the Möbius strip. So we want a developable surface on which c is a geodesic. This can be obtained by taking the rectifying developable of c (Problem 3-13); because of the way c twists, its rectifying developable twists so as to be homeomorphic to the Möbius strip.

Since so many complexities arise even for analytic developables, it might seem hopeless to say anything at all about flat surfaces which are merely C^∞ , and which may contain planar points (in which case we cannot even be sure that they are ruled surfaces). To see how C^∞ flat surfaces can be different from analytic ones, consider the two surfaces pictured below. The first is obtained by



rolling up three pieces of a disc (choosing an appropriate profile for the rolled up portions, so that the surface is C^∞), the second by gluing a cylinder and a cone to a plane. In Chapter 3, our construction of a C^∞ flat Möbius strip gave another example, in which two tangent developables were glued together in a C^∞ , but non-analytic, way. In the above pictures we have singled out certain points p which are planar points, but at the same time the limit of parabolic points. In each case there are segments (indicated by dashed lines) which have p as an endpoint, but which cannot be extended past p in the other direction.

On the other hand, these rays do not go through parabolic points; in fact, most of the points on them lie in completely planar regions. Moreover, in each case there is also a segment (indicated by heavy lines) containing p in its interior. We will show that this situation is completely typical. The main tool is a Lemma which is obtained by following the philosophically prescribed route:

5. LEMMA. Let p be a parabolic point on a flat surface M immersed in \mathbb{R}^3 , let $L_p \subset \mathbb{R}^3$ be the straight line containing the integral curve through p of the vector field of Proposition 4, and let O_p be the component containing p of the set of points in $L_p \cap M$ where $k_2 \neq 0$. Let c be the arclength parameterization of L_p , with $c(0) = p$, and let $k(s) = k_2(c(s))$. Then on O_p the function k is of the form

$$k(s) = \frac{1}{As + B}$$

for some constants A and B .

PROOF. We keep the same notation as in the proof of Proposition 4, so that we have

$$(1) \quad \psi_1^3 = 0, \quad (2) \quad \psi_2^3 = -k_2\theta^2.$$

We have already found, using one of the Codazzi-Mainardi equations, that ω_1^2 is always a multiple of θ^2 . This means that on the region where $k_2 \neq 0$ it is also a multiple of ψ_2^3 , say

$$(3) \quad \omega_1^2 = g\psi_2^3.$$

Now we use the *other* Codazzi-Mainardi equation, to obtain

$$(4) \quad d\psi_2^3 = -\psi_1^3 \wedge \omega_2^1 = 0 \quad \text{by (1).}$$

Thus

$$\begin{aligned} 0 = d\psi_2^3 &= -dk_2 \wedge \theta^2 - k_2 d\theta^2 && \text{by (2)} \\ &= -dk_2 \wedge \theta^2 + k_2 \omega_1^2 \wedge \theta^1, \end{aligned}$$

and therefore

$$dk_2 \wedge \theta^2 = -k_2 \theta^1 \wedge \omega_1^2.$$

Applying this to (X_1, X_2) gives

$$\begin{aligned} (*) \quad X_1(k_2) &= -k_2 \omega_1^2(X_2) \\ &= -k_2 g \psi_2^3(X_2) && \text{by (3)} \\ &= (k_2)^2 g && \text{by (2).} \end{aligned}$$

We also want to use the Gauss equation $d\omega_1^2 = -K\theta^1 \wedge \theta^2$ (page 69). Since $K = 0$, this says that

$$(5) \quad d\omega_1^2 = 0.$$

Hence

$$\begin{aligned} 0 = d\omega_1^2 &= d(g\psi_2^3) && \text{by (3)} \\ &= dg \wedge \psi_2^3 + 0 && \text{by (4)} \\ &= -k_2 dg \wedge \theta^2 && \text{by (2).} \end{aligned}$$

Applying this to (X_1, X_2) gives $dg(X_1) = 0$. In other words,

$$(**) \quad g \text{ is constant along the integral curves of } X_1.$$

From (*) and (**) we see that the function $k(s) = k_2(c(s))$ satisfies the differential equation

$$(***) \quad k'(s) = -Ak(s)^2 \quad \text{for some constant } A.$$

We can solve this explicitly: Since

$$-\frac{k'}{k^2} = \left(\frac{1}{k}\right)',$$

we have

$$k(s) = \frac{1}{As + B}.$$

More precisely, by suitable choice of B we obtain any desired initial condition $k(0)$ except $k(0) = 0$ —the solution with this initial condition is simply $k = 0$. But $k(0) \neq 0$ by assumption, so our k has the above form. ♦

6. COROLLARY. Let p be a parabolic point on a flat surface M immersed in \mathbb{R}^3 . Then there is a unique straight line L_p through p such that the component C_p of $M \cap L_p$ which contains p is an interval (possibly infinite) with p in its interior. All the points of C_p are also parabolic. Moreover, C_p cannot end in M ; that is, if C_p has an endpoint q , then q is not in M .

PROOF. Existence of L_p follows from Proposition 4, and uniqueness is obvious,* since there cannot be two distinct asymptotic directions at p . If there were non-parabolic points of C_p , then one of them would be an endpoint of the interval O_p of Lemma 5. If this point is $q = c(s_0)$, then we would have

$$0 = k_2(q) = k_2(c(s_0)) = \lim_{s \rightarrow s_0} \frac{1}{As + B},$$

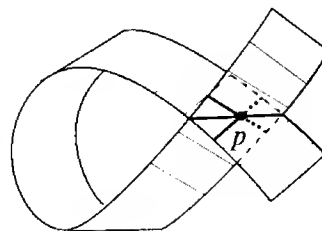
which is impossible. Thus all points of C_p are parabolic, and $C_p = O_p$. Similarly, if C_p had an endpoint $q \in M$, then $k_2(q)$ could not be 0, so q would also be a parabolic point, and a neighborhood of q would be a ruled surface. From this it is clear that C_p could be extended to include q in its interior, which gives a contradiction. ♦

Once we have this Corollary, the next two follow by completely elementary argumentation.

7. COROLLARY. Let M be a flat surface immersed in \mathbb{R}^3 , and let $p \in M$ be a planar point which is a limit of parabolic points p_n . Then the conclusion of Corollary 6 still holds, except that all points of C_p are now planar points which are limits of parabolic points.

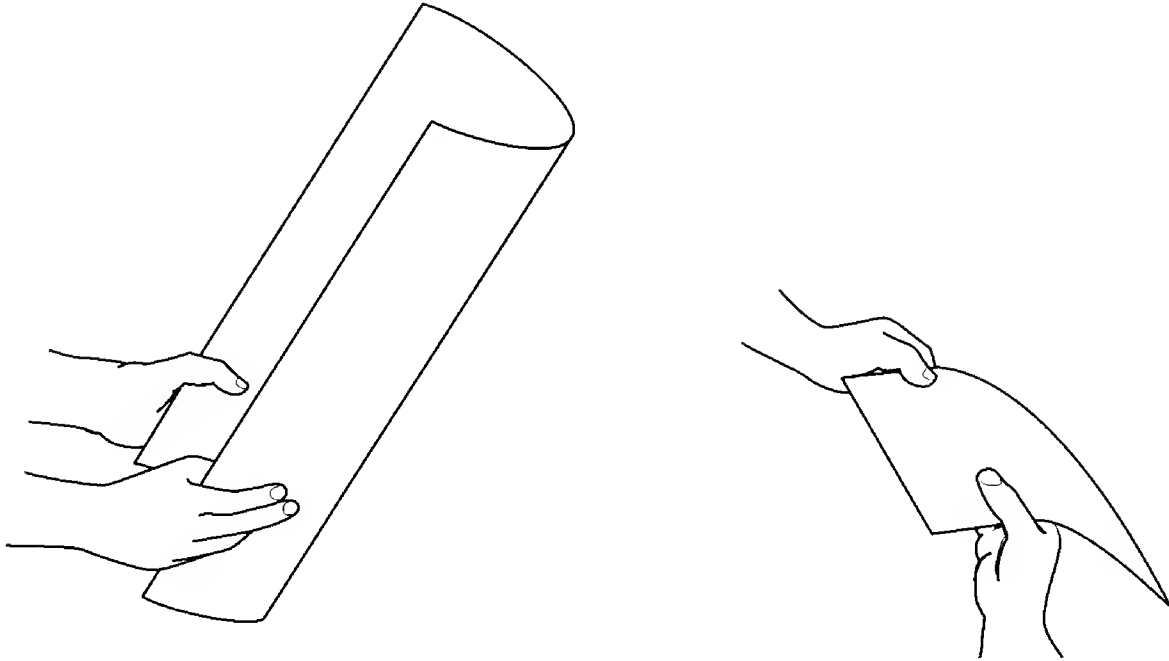
PROOF. Some subsequence of the straight lines L_{p_n} have a limiting direction. Let L_p be the straight line through p with this limiting direction. The components C_{p_n} have p_n in their interior; moreover, Corollary 6 shows that the lengths of the C_{p_n} are bounded away from 0 in each direction from p_n . It follows that C_p has p in its interior. The assertion about endpoints of C_p is an immediate consequence of the same property for the C_{p_n} . It is also clear that all points of C_p are limits of parabolic points, since all points of each C_{p_n} are parabolic. If some point of C_p itself were a parabolic point, then, by Corollary 6, all points of C_p would be parabolic points, including p itself, a contradiction. To prove uniqueness, notice that another straight line L' through p would have to intersect C_{p_n} for large enough n ; thus L' would contain parabolic points, so all points on L' would be parabolic points, including p itself, a contradiction. ♦

*A “counterexample” is shown in the figure on the right: since we are dealing with immersed surfaces, the uniqueness has to be given a careful formulation, which is left to the reader. We will also be somewhat sloppy about such questions in the sequel, since the precise statements are always clear, but irritatingly messy to state.

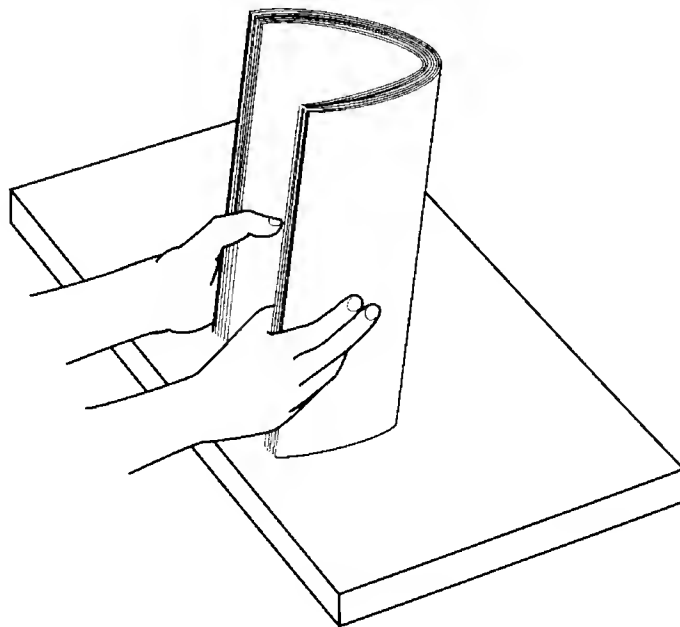


8. COROLLARY. Every point p of a flat surface M immersed in \mathbb{R}^3 is contained in the interior of some line segment lying in M . This segment is unique unless some neighborhood of p is a plane, and the only segments which can end in M are those whose interior points are of this type.

R. Malz likes to point out that this result has a very important application in everyday life. If one holds a piece of paper in the shape of a cylinder, then



it will stay stiff even if it is very long; however, as soon as it is allowed to be planar, it will flop down under gravity. That is why people always curl up a pile of papers when they try to align it by tapping it on a desk top.

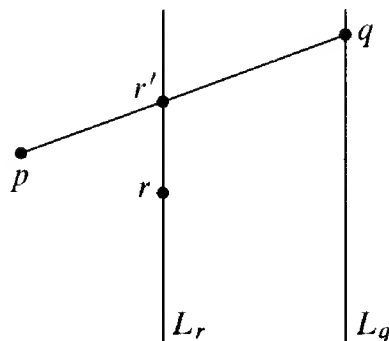


At this point it is almost clear that a *complete* flat surface M immersed in \mathbb{R}^3 must be a generalized cylinder. For an open dense subset of M will be a union of pieces of planes, cylinders, cones and tangent developables, where we can arrange for the latter 3 types to contain only parabolic points. Because of completeness, Corollaries 6 and 7 show that the generators of these cylinders, cones, and tangent developables must be infinite straight lines. But in the case of cones and tangent developables this is simply impossible, for there would definitely be singularities at the vertex or edge of regression. So an open dense subset of M consists of cylinders (some possibly degenerating to planes). It seems fairly clear that these cylinders must all have parallel generators if they and a nowhere dense closed set are somehow going to make up a smooth surface M ; but proving this might become quite sticky. Fortunately, there is a direct proof of the global result which makes no use whatsoever of the classical local classification.

9. THEOREM. If M is a complete flat surface and $f: M \rightarrow \mathbb{R}^3$ is an isometric immersion, then $f(M) \subset \mathbb{R}^3$ is a generalized cylinder.

FIRST PROOF. We can assume that M is simply-connected (by applying the result to $f \circ \pi$ where $\pi: \tilde{M} \rightarrow M$ is the universal covering space). Then (Problem 1-5) M is isometric to \mathbb{R}^2 , with its usual Riemannian metric.

We claim first that if $f(M)$ is not simply a plane in \mathbb{R}^3 , then for every point $p \in M$ there is a *unique* infinite straight line L_p through p which is contained in $f(M)$. Corollaries 6 and 7, together with completeness, show that this is true if p is parabolic or a limit of parabolic points. Now consider a point $p \in f(M)$ which has a whole neighborhood contained in a plane P . Let $Q \subset P$ be the component of $f(M) \cap P$ containing p . If Q is not all of P , let q be a boundary point of Q in P such that all points of the segment \overline{pq} other than q itself lie in the interior of Q (relative to P). The point q must lie in $f(M)$, by completeness,



so q is a planar point which is a limit point of parabolic points. Therefore q is on a unique straight line L_q in $f(M)$. The line L_q must lie in P , since P is the tangent space of q , by continuity of the tangent spaces; moreover, all points of L_q are planar points which are limits of parabolic points.

We claim that all points between p and L_q lie in Q also. Otherwise, there is a point r between p and L_q such that all points of \overline{pr} are in P , but r is also a limit of parabolic points.

There is a corresponding line L_r , and it must be parallel to L_q , since points on L_q and L_r have only one straight line passing through them. Thus L_r intersects \overline{pq} at r' . Then r' is also a limit of parabolic points, which is absurd, since r' is in the interior of Q , a component of $M \cap P$.

The same arguments may be applied if there are any boundary points of Q on the other side of p . Consequently, either $Q = P$, or Q is the part of P bounded by L_q , or Q is the part of P bounded by two parallel lines $L_q, L_{q'}$. Leaving aside the case where $Q = P$ (which occurs only if $f(M) = P$), we see that there is a unique infinite straight line through p in M , namely the one parallel to L_q [and $L_{q'}$]. This proves our claim.

Since f is an isometry, each L_p is the image under f of a geodesic (or possibly many geodesics) in \mathbb{R}^2 ; these geodesics are just ordinary straight lines. In this family of straight lines, distinct lines are disjoint, so our family is the set of lines in \mathbb{R}^2 parallel to a fixed line; for convenience we assume that they are all parallel to the y -axis.

Now by completeness, the functions k of Lemma 5 are defined for all s . But this can happen only if $A = 0$. So we see that all k are constants. In other words,

$$k_2(x, y) = \kappa(x)$$

for some function κ . On the other hand, consider the map $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by

$$g(x, y) = (c_1(x), c_2(x), y),$$

where c is a curve in \mathbb{R}^2 with curvature function κ . We easily compute that the maps f and g both have second fundamental forms with components

$$l = \kappa, \quad m = 0, \quad n = 0.$$

So by the fundamental theorem of surface theory, f differs from the generalized cylinder g by a Euclidean motion.

SECOND PROOF. We replace the last argument, using the fundamental theorem of surface theory, with some very elementary geometry. Any two parallel lines L_1 and L_2 of our family have the property that the function

$$p \mapsto d(p, L_2)$$

is constant on L_1 , where $d(p, L_2)$ is the distance (in \mathbb{R}^2) from p to L_2 . Since f is an isometry, the function $q \mapsto \bar{d}(q, f(L_2))$ must be constant on $f(L_1)$, where \bar{d} denotes the distance in $f(M)$. Now if $f(L_1)$ and $f(L_2)$ were skew lines, or lines intersecting at just one point, then the function

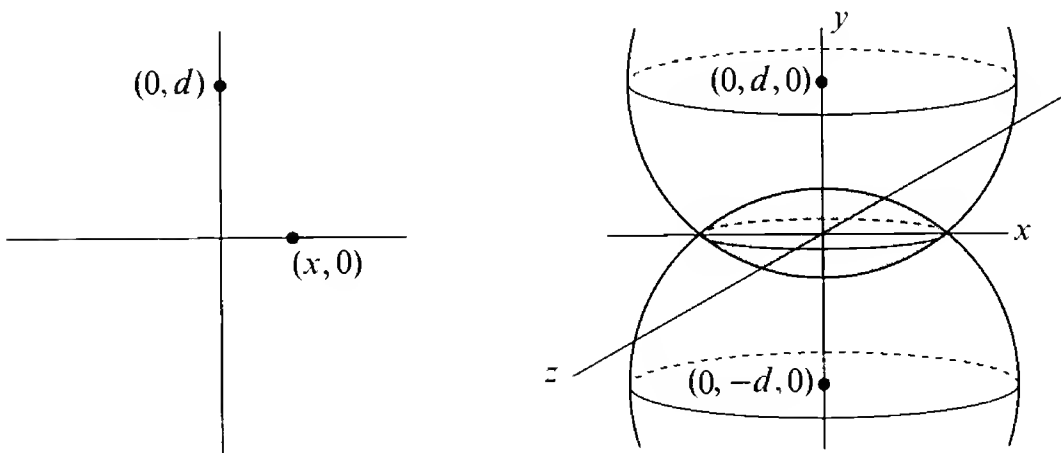
$$q \mapsto \text{Euclidean distance from } q \text{ to } f(L_2)$$

would $\rightarrow \infty$ as $q \rightarrow \infty$ along L_1 . Since

$$\bar{d}(q, f(L_2)) \geq \text{Euclidean distance from } q \text{ to } f(L_2),$$

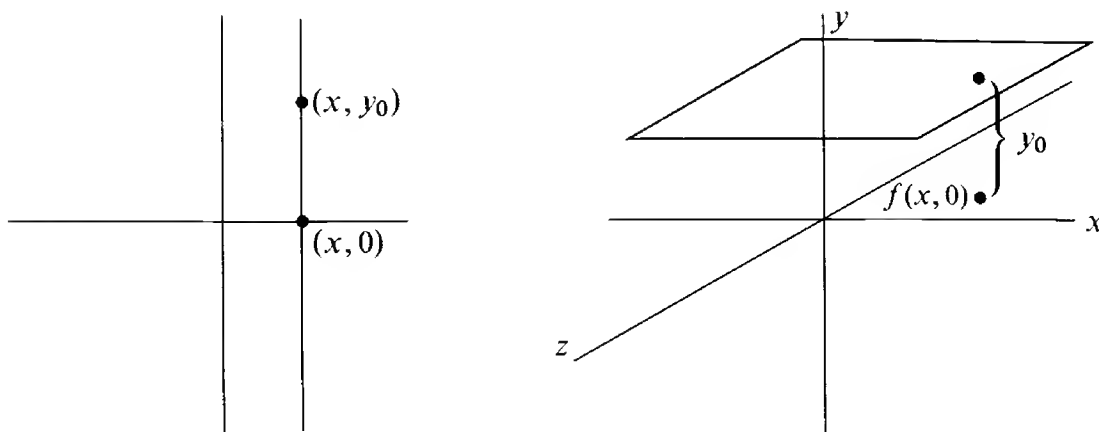
the same would be true for \bar{d} , so \bar{d} could not be constant. Thus $f(L_1)$ and $f(L_2)$ must be parallel (or equal). So $f(M)$ is a generalized cylinder.

THIRD PROOF. This time we reduce almost everything to elementary geometry. We merely note that either all points of our surface are planar points, or by Corollaries 6 and 7, and completeness, *some* straight line of \mathbb{R}^2 maps to a straight line in \mathbb{R}^3 ; for simplicity we assume that $(0, y) \mapsto (0, y, 0)$. Consider first a point $(x, 0)$ of \mathbb{R}^2 . Its distance from $(0, d)$ is $\sqrt{x^2 + d^2}$. Since this must be the distance from $f(x, 0)$ to $f(0, d)$ in $f(M)$, the point $f(x, 0)$ must lie in the Euclidean ball around $(0, d, 0)$ of radius $\sqrt{x^2 + d^2}$. Similarly, $f(x, 0)$ must



lie in the Euclidean ball around $(0, -d, 0)$ of radius $\sqrt{x^2 + d^2}$. The intersection of these two balls is a lens-shaped region of height $2(\sqrt{x^2 + d^2} - d)$. Since this $\rightarrow 0$ as $d \rightarrow \infty$, the point $f(x, 0)$ must lie in the plane $y = 0$.

Now the same argument shows that $f(x, y_0)$ must lie in the plane $y = y_0$.



But $f(x, y_0)$ must also lie in the Euclidean ball of radius y_0 around $f(x, 0)$. So $f(x, y_0)$ must be on the line through $f(x, 0)$ parallel to the y -axis. ♦

The remainder of this chapter is devoted to the proof of Hilbert's theorem that there are *no* complete surfaces of constant negative curvature K immersed in \mathbb{R}^3 . There is no loss of generality in considering only the case $K = -1$, since similarities of \mathbb{R}^3 multiply K by a (positive) constant. We will actually give two proofs of this result. The second is related to, but considerably simpler than, the original proof of Hilbert [1], while the first is an alternative to Hilbert's argument, due to Holmgren [1]. The proofs depend on several classical formulas for surfaces of constant negative curvature, so in each case a few preparatory results are in order.

10. LEMMA. Let M be a 2-dimensional immersed submanifold of \mathbb{R}^3 with constant curvature $K < 0$. Then for every point $p \in M$ there is a diffeomorphism

$$\begin{aligned} g : (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) &\rightarrow M, \\ g(0, 0) &= p, \end{aligned}$$

whose parameter curves are asymptotic curves *parameterized by arclength*.

A CLASSICAL PROOF. In Addendum 1 to Chapter 4 we found that for every $p \in M$ there is a diffeomorphism $g : (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \rightarrow M$, with $g(0, 0) = p$, such that the parameter curves are asymptotic curves. By a suitable reparameterization we can clearly arrange that the two parameter curves through $p = g(0, 0)$ are parameterized by arclength. Thus we have

$$(1) \quad E(s, 0) = 1, \quad G(0, t) = 1.$$

We now claim that *all* parameter curves are parameterized by arclength. To prove this we note that the Codazzi-Mainardi equations on page 217 can be written

$$(2) \quad \begin{aligned} (m^2)_1 &= 2 \frac{\left[\frac{1}{2}(EG - F^2)_1 + FE_2 - EG_1\right]}{EG - F^2} m^2 \\ (m^2)_2 &= 2 \frac{\left[\frac{1}{2}(EG - F^2)_2 + FG_1 - GE_2\right]}{EG - F^2} m^2. \end{aligned}$$

But we also have

$$K = \frac{ln - m^2}{EG - F^2} = \frac{-m^2}{EG - F^2},$$

so that

$$m^2 = (-K)(EG - F^2), \quad \text{where } K \text{ is a constant.}$$

Substituting in the first equation of (2), we get

$$(-K)(EG - F^2)_1 = 2(-K) \left[\frac{1}{2}(EG - F^2)_1 + FE_2 - EG_1 \right],$$

which becomes simply

$$EG_1 - FE_2 = 0.$$

Similarly,

$$-FG_1 + GE_2 = 0.$$

Since $EG - F^2 \neq 0$, this set of linear equations is satisfied only if $E_2 = 0$ and $G_1 = 0$. Together with (1), this shows that $E = 1$ and $G = 1$ everywhere.

SECOND PROOF. The desired result obviously amounts to the following: If Y_1 and Y_2 are linearly independent unit asymptotic vector fields, then $[Y_1, Y_2] = 0$.

First consider an adopted orthonormal moving frame (X_1, X_2, X_3) for which X_1 and X_2 are principal directions, with corresponding principal curvatures k_1 and k_2 . Then the forms ψ_i^3 satisfy

$$\psi_i^3 = k_i \theta^i.$$

We also have the Codazzi-Mainardi equations

$$d\psi_1^3 = \omega_1^2 \wedge \psi_2^3, \quad d\psi_2^3 = -\omega_1^2 \wedge \psi_1^3.$$

Taking the exterior derivative of the equation $\psi_1^3 = k_1 \theta^1$ thus yields

$$dk_1 \wedge \theta^1 + k_1 d\theta^1 = d\psi_1^3 = \omega_1^2 \wedge \psi_2^3 = k_2 \omega_1^2 \wedge \theta^2,$$

so that

$$dk_1 \wedge \theta^1 + k_1 \omega_1^2 \wedge \theta^2 = k_2 \omega_1^2 \wedge \theta^2.$$

Applying this to (X_1, X_2) , we find that

$$-X_2(k_1) = (k_2 - k_1) \cdot \omega_1^2(X_1).$$

Similarly, exterior differentiation of the equation $\psi_2^3 = k_2 \theta^2$ leads to

$$-X_1(k_2) = (k_2 - k_1) \cdot \omega_1^2(X_2).$$

Thus

$$(1) \quad \omega_1^2 = -\frac{X_2(k_1)}{k_2 - k_1} \theta^1 - \frac{X_1(k_2)}{k_2 - k_1} \theta^2.$$

Now consider a unit vector field

$$\alpha_1 X_1 + \alpha_2 X_2,$$

with

$$(2) \quad (\alpha_1)^2 + (\alpha_2)^2 = 1.$$

For this to be an asymptotic vector field we need

$$k_1(\alpha_1)^2 + k_2(\alpha_2)^2 = 0.$$

For simplicity we now take the case $K = -1$, so that $k_1 k_2 = -1$. Then $\alpha_1 X_1 + \alpha_2 X_2$ is asymptotic if and only if

$$\begin{aligned} (\alpha_2)^2 = -\frac{k_1}{k_2}(\alpha_1)^2 &\implies (\alpha_1)^2 - \frac{k_1}{k_2}(\alpha_1)^2 = 1 \quad \text{by (2)} \\ &\implies (\alpha_1)^2 = \frac{1}{1 - \frac{k_1}{k_2}} \\ &= \frac{1}{1 + k_1^2}, \quad \text{since } k_1 k_2 = -1. \end{aligned}$$

So our unit asymptotic vector fields must be

$$\begin{aligned} Y_1 &= \pm \alpha_1 X_1 \pm \alpha_2 X_2 \\ Y_2 &= \pm \alpha_1 X_1 \pm \alpha_2 X_2, \end{aligned}$$

where the α_i are given by

$$(3) \quad \alpha_i = \frac{1}{\sqrt{1 + k_i^2}}.$$

It is convenient to note that we can write equation (1) in terms of the α_i as

$$(4) \quad \omega_j^i = \frac{X_j(\alpha_i)}{\alpha_i} \theta^i - \frac{X_i(\alpha_j)}{\alpha_j} \theta^j.$$

Now to show that $[Y_1, Y_2] = 0$ we just need to show that $[\alpha_1 X_1, \alpha_2 X_2] = 0$. But we have (see pg. I.215)

$$\begin{aligned} \theta^i([\alpha_1 X_1, \alpha_2 X_2]) &= \alpha_1 X_1(\theta^i(\alpha_2 X_2)) - \alpha_2 X_2(\theta^i(\alpha_1 X_1)) - d\theta^i(\alpha_1 X_1, \alpha_2 X_2) \\ &= \delta_{i2} \alpha_1 X_1(\alpha_2) - \delta_{i1} \alpha_2 X_2(\alpha_1) + \sum_j (\omega_j^i \wedge \theta^j)(\alpha_1 X_1, \alpha_2 X_2) \\ &= \delta_{i2} \alpha_1 X_1(\alpha_2) - \delta_{i1} \alpha_2 X_2(\alpha_1) \\ &\quad + \delta_{i1} \alpha_2 X_2(\alpha_1) - \delta_{i2} \alpha_1 X_1(\alpha_2) \quad \text{using (4)} \\ &= 0. \quad \blacklozenge \end{aligned}$$

For any 2-dimensional Riemannian manifold M (not necessarily immersed in \mathbb{R}^3), an immersion $g: (a, b) \times (c, d) \rightarrow M$ is called a **Tschebyscheff net** if all parameter curves are parameterized by arclength. If we think of the domain $(a, b) \times (c, d)$ as a piece of cloth woven from fibres parallel to the axes, then the immersion g doesn't stretch any fibres. So the surface can be outfitted in a sexy tight fitting suit if we can find Tschebyscheff nets around each point (we might have to sew a lot of pieces together). Lemma 10 shows that this can always be done on a submanifold of \mathbb{R}^3 with constant negative curvature. The notion of a Tschebyscheff net is an intrinsic one, however, and our next result is also.

11. LEMMA. Let M be a 2-dimensional Riemannian manifold and $g: (a, b) \times (c, d) \rightarrow M$ a Tschebyscheff net. Define $\omega: (a, b) \times (c, d) \rightarrow \mathbb{R}$ as follows: $\omega(s_0, t_0)$ is the unique number with $0 < \omega(s_0, t_0) < \pi$ such that $\omega(s_0, t_0)$ is an angle between

$$\left. \frac{dg(s, t_0)}{ds} \right|_{s=s_0} \quad \text{and} \quad \left. \frac{dg(s_0, t)}{dt} \right|_{t=t_0}$$

Then ω satisfies the differential equation

$$\frac{\partial^2 \omega}{\partial s \partial t} = (-K) \sin \omega.$$

PROOF. We have $E = G = 1$, and

$$F = \cos \omega, \quad W = \sqrt{EG - F^2} = \sin \omega.$$

From the equation in Problem 4-13 we obtain

$$\begin{aligned} K &= \frac{1}{2W} \left[\frac{\partial}{\partial t} \left(\frac{F_1}{W} \right) + \frac{\partial}{\partial s} \left(\frac{F_2}{W} \right) \right] \\ &= \frac{1}{2W} \left[\frac{\partial}{\partial t} \left(\frac{-\sin \omega \frac{\partial \omega}{\partial s}}{W} \right) + \frac{\partial}{\partial s} \left(\frac{-\sin \omega \frac{\partial \omega}{\partial t}}{W} \right) \right] \\ &= \frac{1}{2 \sin \omega} \left[\frac{\partial}{\partial t} \left(-\frac{\partial \omega}{\partial s} \right) + \frac{\partial}{\partial s} \left(-\frac{\partial \omega}{\partial t} \right) \right] \\ &= -\frac{\frac{\partial^2 \omega}{\partial s \partial t}}{\sin \omega}. \quad \diamond \end{aligned}$$

We are now ready to prove the theorem, which still requires quite a bit of argument. We will use the term **asymptotic Tschebyscheff net** for a Tschebyscheff net of the sort constructed in Lemma 10, with all parameter curves being asymptotic curves.

12. THEOREM. A complete surface M with constant curvature $K = -1$ cannot be immersed in \mathbb{R}^3 .

PROOF. The proof depends on establishing two facts:

- (A) Suppose that M could be immersed in \mathbb{R}^3 . Then there would be a Tschebyscheff net $f: \mathbb{R}^2 \rightarrow M$, from the whole plane to M , and the function ω , defined on all of \mathbb{R}^2 , which gives the angle between the first and second parameter lines would satisfy

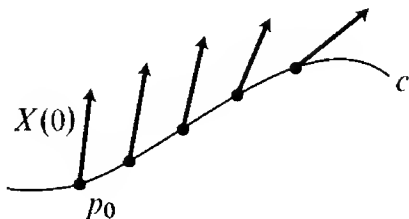
$$\frac{\partial^2 \omega}{\partial s \partial t} = \sin \omega, \quad 0 < \omega < \pi.$$

- (B) There is no function $\omega: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfying

$$\frac{\partial^2 \omega}{\partial s \partial t} = C \sin \omega, \quad 0 < \omega < \pi,$$

where $C > 0$ is any constant.

PROOF OF (A). Select a point $p_0 \in M$. Let $c: \mathbb{R} \rightarrow M$ be an asymptotic curve, parameterized by arclength, with $c(0) = p_0$; since c is locally an integral curve of a *unit* vector field, it can be defined on all of \mathbb{R} since M is complete [compare Problem 1-5(c)]. Let $X(0)$ be a unit asymptotic vector at p_0 which is linearly independent of $c'(0)$, and let $X(t)$ be the unique continuous vector field X along c such that $X(t) \in M_{c(t)}$ is a unit asymptotic vector linearly independent



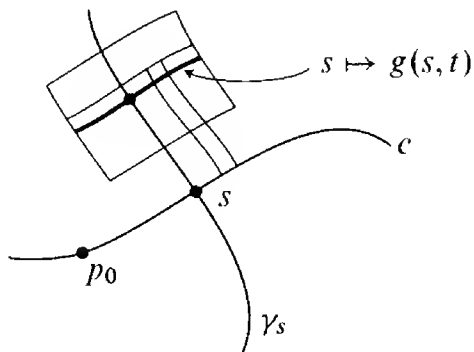
of $c'(t)$. The vector field X along c is just a device to enable us to distinguish a direction for each parameter value of c . We now define $f: \mathbb{R}^2 \rightarrow M$ as follows:

$$f(s, t) = \gamma_s(t),$$

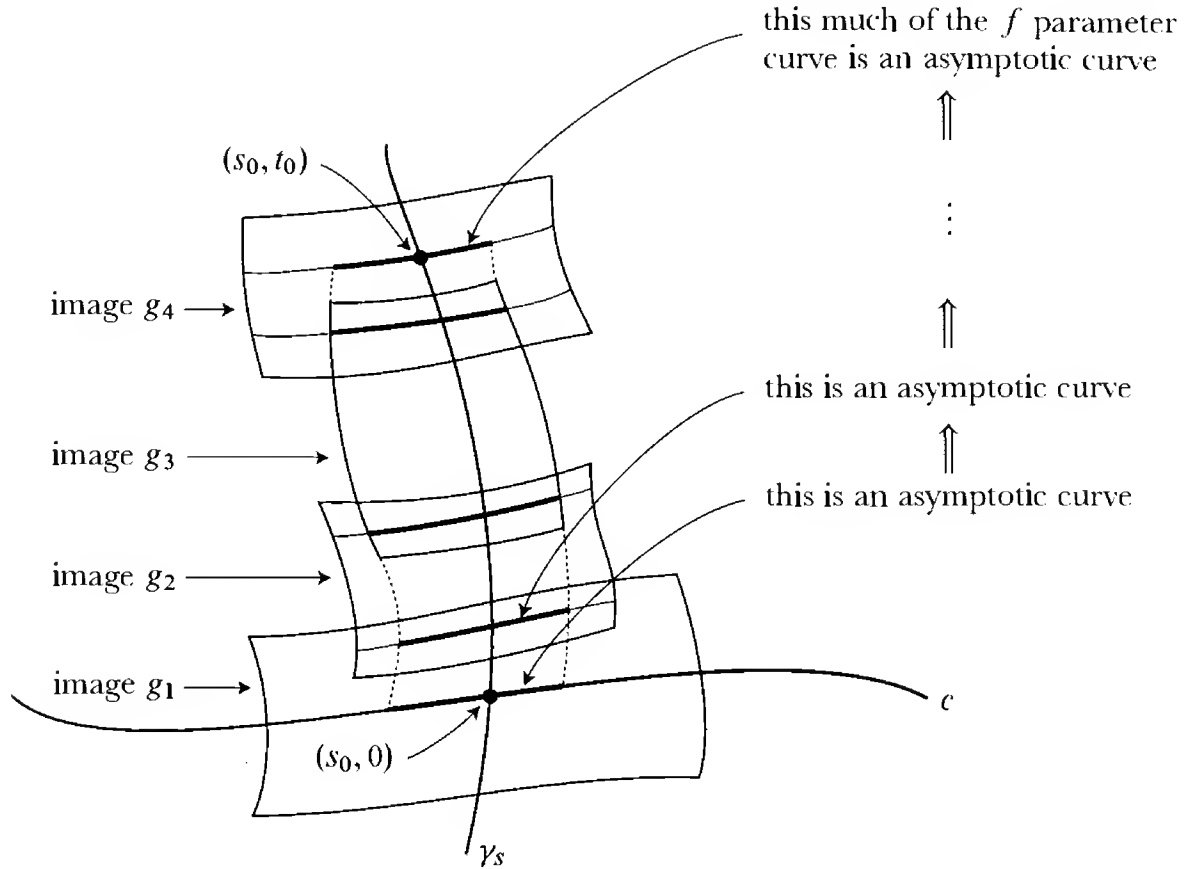
where γ_s is the unique asymptotic curve, parameterized by arclength, with $\gamma_s(0) = c(s)$ and $\gamma_s'(0) = X(s)$.

What we have to show is that each curve $s \mapsto f(s, t)$ is an asymptotic curve. This depends on the existence, as guaranteed by Lemma 10, of asymptotic Tschebyscheff nets $g: (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \rightarrow M$ around any point. From the very definition of f the following at least is clear:

Observation: If for some $t \in (-\varepsilon, \varepsilon)$ the parameter curve $s \mapsto g(s, t)$ lies along a parameter curve $s \mapsto f(s, \bar{t})$, then *all* parameter curves $s \mapsto g(s, t)$ lie along parameter curves of f .



Now for any (s_0, t_0) we can find a finite number of asymptotic Tschebyscheff nets g_1, \dots, g_k whose images cover $\{f(s_0, t): 0 \leq t \leq t_0\}$. Arranging the g_i as in the picture below, so that the images of consecutive ones overlap, noting that



$s \mapsto f(s, 0)$ is an asymptotic curve by definition, and applying the Observation repeatedly, we see that $s \mapsto f(s, t_0)$ is an asymptotic curve for s sufficiently close to s_0 , which is what we wanted.

Our equation for ω then follows immediately from Lemma 11.

PROOF OF (B). Suppose we had a function $\omega: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfying

$$(1) \quad \frac{\partial^2 \omega}{\partial s \partial t} = C \sin \omega, \quad 0 < \omega < \pi,$$

for a constant $C > 0$, and hence, in particular,

$$(2) \quad \frac{\partial^2 \omega}{\partial s \partial t} > 0.$$

This implies that $\partial \omega / \partial s$ is increasing as a function of t , so that

$$(3) \quad \frac{\partial \omega}{\partial s}(s, t) > \frac{\partial \omega}{\partial s}(s, 0) \quad \text{for } t > 0.$$

Consequently, for $t > 0$ we have

$$\int_a^b \frac{\partial \omega}{\partial s}(s, t) ds > \int_a^b \frac{\partial \omega}{\partial s}(s, 0) ds,$$

so that

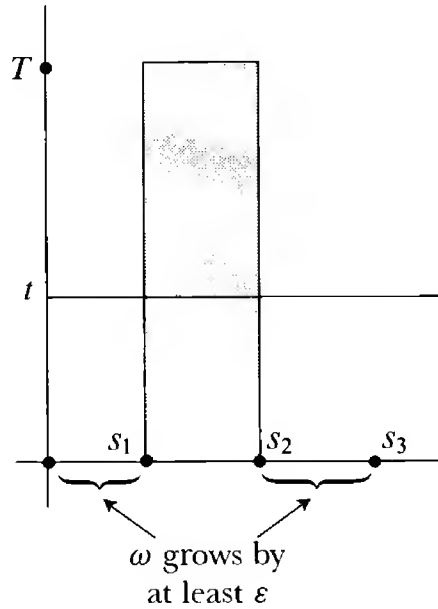
$$(4) \quad \omega(b, t) - \omega(a, t) > \omega(b, 0) - \omega(a, 0) \quad \text{for } t > 0 \text{ and } a < b.$$

Now we can't have $\partial \omega / \partial s = 0$ everywhere, so we can assume (changing our coordinates by a translation) that $\partial \omega / \partial s(0, 0) \neq 0$. Since the function $(s, t) \mapsto \omega(-s, -t)$ also satisfies (1), we can even assume that $\partial \omega / \partial s(0, 0) > 0$. Choose three fixed numbers

$$(5) \quad 0 < s_1 < s_2 < s_3 \quad \text{with} \quad \frac{\partial \omega}{\partial s}(s, 0) > 0 \quad \text{for } 0 \leq s \leq s_3,$$

and let

$$\varepsilon = \min \begin{cases} \omega(s_3, 0) - \omega(s_2, 0) \\ \omega(s_1, 0) - \omega(0, 0). \end{cases}$$



Then for all $t > 0$ and all $s \in [0, s_3]$ we have the following:

$$\left. \begin{array}{l} \omega(s, t) \text{ is increasing in } s, \\ \omega(s_1, t) - \omega(0, t) > \varepsilon \\ \omega(s_3, t) - \omega(s_2, t) > \varepsilon \end{array} \right\}, \quad \begin{array}{l} \text{by (3) and (5)} \\ \text{by (4) and the definition of } \varepsilon \end{array}$$

$$0 < \omega(s, t) < \pi.$$

Putting these together, we conclude that

$$\varepsilon \leq \omega(s, t) \leq \pi - \varepsilon \quad \text{for } s \in [s_1, s_2] \text{ and } t \geq 0,$$

and hence

$$(6) \quad \sin \omega(s, t) \geq \sin \varepsilon \quad \text{for } s \in [s_1, s_2] \text{ and } t \geq 0.$$

But suppose we integrate equation (1) over the rectangle $[s_1, s_2] \times [0, T]$. We obtain

$$\begin{aligned} C \int_0^T \int_{s_1}^{s_2} \sin \omega(s, t) ds dt &= \int_0^T \int_{s_1}^{s_2} \frac{\partial^2 \omega}{\partial s \partial t} ds dt \\ &= [\omega(s_2, T) - \omega(s_1, T) - \omega(s_2, 0) + \omega(s_1, 0)], \end{aligned}$$

or

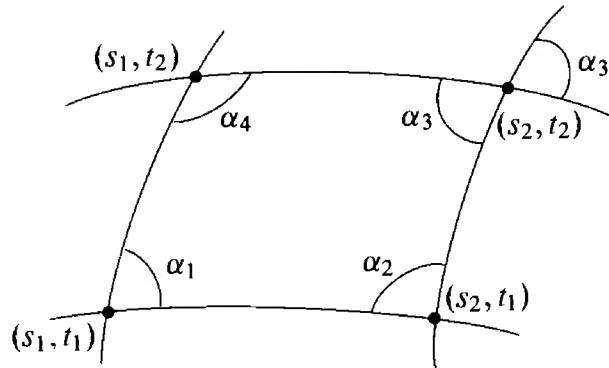
$$\begin{aligned} \omega(s_2, T) - \omega(s_1, T) &= \omega(s_2, 0) - \omega(s_1, 0) + C \int_0^T \int_{s_1}^{s_2} \sin \omega(s, t) ds dt \\ &\geq \omega(s_2, 0) - \omega(s_1, 0) + CT(s_2 - s_1) \sin \varepsilon, \quad \text{by (6).} \end{aligned}$$

Taking T large enough, we get a contradiction, since the left side is $< \pi$. ♦

For the second proof of Hilbert's theorem, we need an observation which follows directly from Lemma 11.

13. LEMMA (HAZZIDAKIS' FORMULA). Let M be a 2-dimensional Riemannian manifold of constant curvature $K < 0$, and let $g: (a, b) \times (c, d) \rightarrow M$ be a Tschebyscheff net. Then any quadrilateral Q formed by parameter curves has area

$$\begin{aligned} \text{area}(Q) &= \frac{1}{-K} \left(\sum_{i=1}^4 \alpha_i - 2\pi \right) \\ &\leq \frac{2\pi}{-K}, \end{aligned}$$



where $\alpha_i \in (0, \pi)$ are the interior angles of Q .

PROOF. Introducing the function ω of Lemma 11, we have

$$dA = W ds \wedge dt = \sin \omega ds \wedge dt$$

$$\sin \omega = \frac{1}{-K} \frac{\partial^2 \omega}{\partial s \partial t}.$$

So

$$\begin{aligned} \text{area}(Q) &= \int_Q dA = \int_Q \sin \omega ds \wedge dt \\ &= \frac{1}{-K} \int_Q \frac{\partial^2 \omega}{\partial s \partial t} ds \wedge dt \\ &= \frac{1}{-K} \int_{s_1}^{s_2} \int_{t_1}^{t_2} \frac{\partial^2 \omega}{\partial s \partial t} ds dt \\ &= \frac{1}{-K} [\omega(s_2, t_2) - \omega(s_1, t_2) - \omega(s_2, t_1) + \omega(s_1, t_1)] \\ &= \frac{1}{-K} [\alpha_3 - (\pi - \alpha_4) - (\pi - \alpha_2) + \alpha_1] \\ &= \frac{1}{-K} \left(\sum_{i=1}^4 \alpha_i - 2\pi \right). \diamond \end{aligned}$$

On the other hand, consider the upper half-plane $\mathcal{H}^2 \subset \mathbb{R}^2$ with the Riemannian metric

$$\langle \cdot, \cdot \rangle = \frac{dx \otimes dx + dy \otimes dy}{y^2}.$$

This is a complete 2-dimensional Riemannian manifold of constant curvature $K = -1$ (see pg. II.301 and Problem I.9-41). We have

$$E = G = \frac{1}{y^2} \quad F = 0$$

$$dA = \sqrt{EG - F^2} dx \wedge dy = \frac{1}{y^2} dx \wedge dy,$$

so the total area of \mathcal{H}^2 is

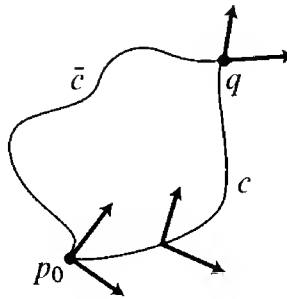
$$\int_{-\infty}^{\infty} \int_0^{\infty} \frac{1}{y^2} dy dx = \infty.$$

Thus Lemma 13 shows that we *cannot* parameterize all of \mathcal{H}^2 by a Tschebyscheff net. This is the basis of our second proof of

12. THEOREM. A complete surface M of constant curvature $K = -1$ cannot be immersed in \mathbb{R}^3 .

PROOF. As in the proof of Theorem 9, we can assume that M is simply-connected. Then (Problem 1-5) M is isometric to $(\mathcal{H}^2, \langle \cdot, \cdot \rangle)$.

We claim first that there are two linearly independent unit asymptotic vector fields Y_1, Y_2 defined *on all of* M . The proof uses the fact that M is simply-connected, and follows a standard procedure. We arbitrarily select $(Y_1(p_0), Y_2(p_0))$ for some $p_0 \in M$. Then for every curve $c: [0, 1] \rightarrow M$ with $c(0) = p_0$, there will be a unique possible continuous choice of $(Y_1(c(t)), Y_2(c(t)))$ for all $t \in [0, 1]$ which extends $(Y_1(p_0), Y_2(p_0))$. If \bar{c} is another such curve with corresponding



$(\bar{Y}_1(c(t)), \bar{Y}_2(c(t)))$, and if moreover $\bar{c}(1) = c(1) = q$, then $\bar{Y}_i(q) = Y_i(q)$; the proof uses the usual argument involving a contraction to the constant path at p_0 of the curve c followed by \bar{c} in the reverse direction. A nicer argument is the following. For each p there are 4 possible choices for $Y_1(p)$, and then 2 possible choices for $Y_2(p)$, thus 8 possible choices for $(Y_1(p), Y_2(p))$. The set of all such pairs, for all $p \in M$, obviously forms an 8-fold covering space of M . Since M is simply-connected, this covering space consists of 8 components; any component gives us the desired pair of vector fields.

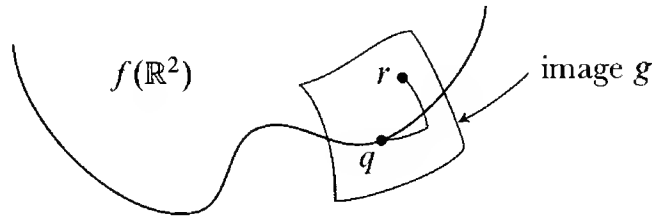
We now claim that the Tschebyscheff net $f: \mathbb{R}^2 \rightarrow M$ which we constructed in the first proof is actually a *diffeomorphism*. To prove this, we first describe f a little differently. Let $\{\phi_t^i\}$ be the 1-parameter group of diffeomorphisms generated by Y_i (recall this means that $t \mapsto \phi_t^i(p)$ is the integral curve of Y_i through p , for each p); the ϕ_t^i can be defined for all $t \in \mathbb{R}$ since M is complete and Y_i are unit vector fields (as in our first proof). We now pick a point $p_0 \in M$ and define

$$(1) \quad f(s, t) = \phi_t^2(\phi_s^1(p_0)).$$

Since $[Y_1, Y_2] = 0$ (Lemma 10), the 1-parameter groups ϕ^1, ϕ^2 commute (Lemma I.5-13), so we easily find that

$$(2) \quad f(s + s', t + t') = \phi_{t'}^2(\phi_{s'}^1(f(s, t))).$$

We claim first that f is onto M . Otherwise, there is a point $q \notin f(\mathbb{R}^2)$ with q on the boundary of $f(\mathbb{R}^2)$. Now there is an asymptotic Tschebyscheff net $g: (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \rightarrow M$ with $g(0, 0) = q$, and there is some point $r \in$



$(\text{image } g) \cap f(\mathbb{R}^2)$. Then q must be of the form

$$\begin{aligned} q &= \phi_{t'}^2(\phi_{s'}^1(r)) && \text{for some } s', t' \\ &= \phi_{t'}^2(\phi_{s'}^1(f(s, t))) && \text{for some } s, t \\ &= f(s + s', t + t') && \text{by (2),} \end{aligned}$$

so actually $q \in f(\mathbb{R}^2)$, a contradiction.

Now we claim that $f: \mathbb{R}^2 \rightarrow M$ is actually a covering map. For any point $q \in M$, choose an asymptotic Tschebyscheff net $g: (-2\varepsilon, 2\varepsilon) \times (-2\varepsilon, 2\varepsilon) \rightarrow M$ with $g(0, 0) = q$ such that g is a diffeomorphism onto some open subset $V \subset M$. We can assume that the s [and t] parameter curves of g lie along the s [and t] parameter curves of f . Suppose that $(s, t) \in f^{-1}(q)$. Consider the map

$$\phi: V \rightarrow (s - 2\varepsilon, s + 2\varepsilon) \times (t - 2\varepsilon, t + 2\varepsilon) \subset \mathbb{R}^2$$

defined by

$$g(s', t') \mapsto (s + s', t + t').$$



Using equation (2), we see that $f: (s - 2\varepsilon, s + 2\varepsilon) \times (t - 2\varepsilon, t + 2\varepsilon) \rightarrow V$ is a diffeomorphism with inverse ϕ . Now let

$$W = g((-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon))$$

and for each $(s, t) \in \mathbb{R}^2$ let

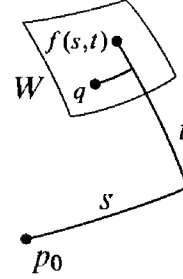
$$W_{(s,t)} = (s - \varepsilon, s + \varepsilon) \times (t - \varepsilon, t + \varepsilon).$$

We claim that

$$f^{-1}(W) = \bigcup_{(s,t) \in f^{-1}(q)} W_{(s,t)}.$$

In fact, if

$$(s, t) \in f^{-1}(W),$$



then we have

$$f(s, t) = g(s', t') \quad \text{for some } (s', t') \in (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon),$$

and by equation (2),

$$\begin{aligned} f(s - s', t - t') &= \phi_{-t'}^2(\phi_{-s'}^1(f(s, t))) \\ &= \phi_{-t'}^2(\phi_{-s'}^1(g(s', t'))) \\ &= q, \end{aligned}$$

which proves the claim. Since each of the $W_{(s,t)}$ is mapped diffeomorphically onto W , the proof that f is a covering map will be complete once we show that any two distinct such rectangles $W_{(s_1, t_1)}$ and $W_{(s_2, t_2)}$ are disjoint. Now if $W_{(s_1, t_1)} \cap W_{(s_2, t_2)} \neq \emptyset$, then

$$(s_2, t_2) \in (s_1 - 2\varepsilon, s_1 + 2\varepsilon) \times (t_1 - 2\varepsilon, t_1 + 2\varepsilon).$$

But we know that f is a diffeomorphism on this rectangle. Since $f(s_1, t_1) = q = f(s_2, t_2)$, this means that $(s_1, t_1) = (s_2, t_2)$, so that $W_{(s_1, t_1)}$ and $W_{(s_2, t_2)}$ are actually the same. Thus f is indeed a covering map.

Now by the simple-connectivity of M we conclude that f is actually a diffeomorphism. Consequently, we can exhaust M by quadrilaterals formed by the parameter curves of f . Lemma 13 then implies that M has area $\leq 2\pi$, while we computed that M has infinite area. This contradiction establishes the theorem. ♦

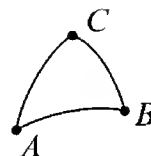
It is easy to find a complete surface $M \subset \mathbb{R}^3$ with non-constant curvature $K < 0$ everywhere—for example, the elliptic hyperboloid of one sheet has this property. But in this example K comes arbitrarily close to 0. In 1964, Efimov [1] proved, by a lengthy ingenious argument, using no particularly sophisticated machinery, that there are no complete surfaces $M \subset \mathbb{R}^3$ with curvature $K < 0$ bounded away from 0; an exposition of the proof may be found in Klotz [2].

CHAPTER 6

THE GAUSS-BONNET THEOREM AND RELATED TOPICS

In Volume II we presented Gauss' proof that if $\triangle ABC$ is a geodesic triangle on a surface with Gaussian curvature K , then

$$\int_{\triangle ABC} K dA = \angle A + \angle B + \angle C - \pi.$$

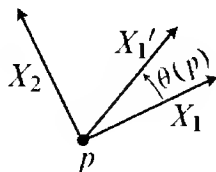


At that time we also cast a suspicious glance at Stokes' Theorem, which seemed to be lurking in the background, and promised to present a proof which would implicate it more fully. We are now in a position to redeem that pledge. It is almost a foregone conclusion that moving frames will play a leading role in the proceedings, since this is the only treatment of curvature in which differential forms appear explicitly. Actually, we are going to generalize Gauss' result, and in two quite different directions. On the one hand, we will allow polygons with any number of sides, and we will not require the sides to be geodesics. On the other hand, we will also have something to say about the integral of K over a whole surface. We begin much more modestly, however, with quite elementary considerations.

Suppose that $\mathbf{X} = (X_1, X_2)$ and $\mathbf{X}' = (X'_1, X'_2)$ are two orthonormal moving frames on a 2-dimensional Riemannian manifold M . We have already found the relationship between the matrices of 1-forms $\omega = (\omega_j^i)$ and $\omega' = (\omega_j'^i)$ associated to these moving frames: If $\mathbf{X}' = \mathbf{X} \cdot a$ for an orthonormal matrix function a , then (see pg. II.280) we have

$$(I) \quad \omega' = a^{-1} da + a^{-1} \omega a.$$

Of course, in a 2-dimensional manifold this relationship can be expressed much more simply. If $\mathbf{X}(p)$ and $\mathbf{X}'(p)$ are similarly oriented, then the matrix $a(p)$



is just

$$a(p) = \begin{pmatrix} \cos \theta(p) & \sin \theta(p) \\ -\sin \theta(p) & \cos \theta(p) \end{pmatrix},$$

where $\theta(p)$ is the oriented angle between $X_1(p)$ and $X'_1(p)$.

Usually we define the “angle” between two vectors to be a number between $-\pi$ and π , but then the function θ need not be continuous. Locally, we can make θ differentiable by allowing other values of θ . In the next result, it does not matter that this θ is not well-defined, because the form $d\theta$ still is.

1. PROPOSITION. Let X_1, X_2 and X'_1, X'_2 be two similarly oriented orthonormal moving frames on a 2-dimensional Riemannian manifold M , and let $\omega_1^2, \omega_1'^2$ be the corresponding connection forms. Let θ be a differentiable choice of the angle between X_1 and X'_1 . Then

$$\omega_1'^2 = \omega_1^2 + d\theta.$$

PROOF. It is easily checked that this is precisely what equation (1) comes down to. It is probably also a good exercise to derive the whole thing from scratch, using properties of ∇ , or by reproving Proposition II.7-14 for 2-manifolds. ♦

Now consider a curve $c: [a, b] \rightarrow M$ which lies in a region on which we have an orthonormal moving frame X_1, X_2 . Suppose that V is a unit vector field along c . We can then define the angle between V and X_1 in an even more precise manner, based on the constructions on pp. II.16–18. We first define a map $\alpha: [a, b] \rightarrow S^1$ by letting $\alpha(t)$ be the image of $V(t)$ under the unique linear map $M_{c(t)} \rightarrow \mathbb{R}^2$ which takes $X_1(c(t))$ to e_1 . We then have, by



Proposition II.1-5, a continuous map $\phi: [a, b] \rightarrow \mathbb{R}$ with

$$\alpha(t) = (\cos \phi(t), \sin \phi(t)),$$

and any two such maps differ by a multiple of 2π . We will refer to any such ϕ as a continuous choice of the angle between X_1 and V . It is easy to see that ϕ is actually C^∞ if V is a C^∞ vector field along c .

2. COROLLARY. Let M be a 2-dimensional Riemannian manifold, and let $c: [a, b] \rightarrow M$ be an immersed curve which lies in a region on which we have a positively oriented orthonormal moving frame X_1, X_2 . Let V be a C^∞ unit vector field along c , and let ϕ be a continuous choice of the angle between X_1 and V . Then V is parallel along c if and only if

$$\omega_1^2(c'(t)) + \phi'(t) = 0.$$

PROOF. Locally we can find an orthonormal moving frame X'_1, X'_2 oriented similarly to X_1, X_2 , with $X'_1 = V$ along c . We can choose the angle $\theta(p)$ between $X'_1(p)$ and $X_1(p)$ so that

$$\theta(c(t)) = \phi(t).$$

Then V is parallel along c if and only if

$$\begin{aligned} 0 &= \langle \nabla_{c'(t)} X'_1, X'_2 \rangle = \omega_1'^2(c'(t)) \\ &= \omega_1^2(c'(t)) + d\theta(c'(t)) \quad \text{by Proposition 1} \\ &= \omega_1^2(c'(t)) + (\theta \circ c)'(t) \\ &= \omega_1^2(c'(t)) + \phi'(t). \quad \spadesuit \end{aligned}$$

In Chapter 4 we defined the geodesic curvature κ_g of a curve c in an oriented surface $M \subset \mathbb{R}^3$. We also noted that κ_g is intrinsic. Indeed, for any arclength parameterized curve c in any Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ we can define κ_g (≥ 0) as the norm of $Dc'(s)/ds$; if M is an oriented 2-dimensional Riemannian manifold, then we defined the signed geodesic curvature κ_g by

$$\kappa_g(s) = \left\langle \frac{Dc'(s)}{ds}, \mathbf{u}(s) \right\rangle,$$

where $\mathbf{u}(s) \in M_{c(s)}$ is the unit vector perpendicular to $c'(s)$ with $(c'(s), \mathbf{u}(s))$ positively oriented.

3. COROLLARY. Let M be an oriented 2-dimensional Riemannian manifold, and let $c: [a, b] \rightarrow M$ be a curve, parameterized by arclength, which lies in a region on which we have a positively oriented orthonormal moving frame X_1, X_2 . Let ϕ be a continuous choice of the angle between X_1 and $c'(s)$. Then the signed geodesic curvature κ_g of c is given by

$$\kappa_g(s) = \omega_1^2(c'(s)) + \phi'(s).$$

PROOF. Locally we can find a positively oriented orthonormal moving frame X'_1, X'_2 with $X'_1 = c'$ along c . So we can choose the angle $\theta(p)$ between $X'_1(p)$ and $X_1(p)$ so that

$$\theta(c(s)) = \phi(s).$$

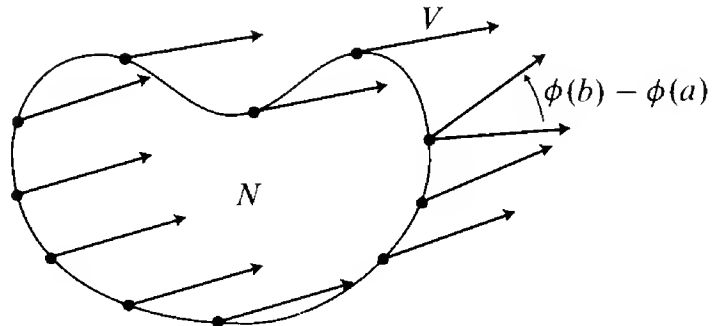
Then

$$\begin{aligned} \kappa_g(s) &= \langle \nabla_{X'_1} X'_1, X'_2 \rangle(c(s)) = \omega_1'^2(c'(s)) \\ &= \omega_1^2(c'(s)) + d\theta(c'(s)) \quad \text{by Proposition 1} \\ &= \omega_1^2(c'(s)) + \phi'(s). \quad \blacklozenge \end{aligned}$$

Each of our Corollaries can be used to obtain an interesting result. The first theorem partially fulfills a promise made on pg. II.243, for it gives a quantitative description of the fact that parallel translation along a closed curve generally does not bring a vector back to itself.

4. THEOREM. Let M be an oriented 2-dimensional Riemannian manifold, with Gaussian curvature K , and volume element dA . Let $N \subset M$ be a compact 2-dimensional manifold-with-boundary whose boundary ∂N is connected, let $c: [a, b] \rightarrow \partial N$ be a closed curve such that $c'(t)$ is positively oriented (with respect to the induced orientation on ∂N), and let V be a parallel unit vector field along c . If X_1, X_2 is a positively oriented moving frame defined on N , and $\phi: [a, b] \rightarrow \mathbb{R}$ is a continuous choice of the angle between X_1 and V , then

$$\phi(b) - \phi(a) = \int_N K dA.$$



PROOF. By the equations on page 69 we have

$$\begin{aligned}
 \int_N K dA &= \int_N K \theta^1 \wedge \theta^2 = - \int_N d\omega_1^2 \\
 &= - \int_{\partial N} \omega_1^2 \quad \text{by Stokes' Theorem} \\
 &= - \int_a^b \omega_1^2(c'(t)) dt \\
 &= \int_a^b \phi'(t) dt \quad \text{by Corollary 2} \\
 &= \phi(b) - \phi(a). \quad \spadesuit
 \end{aligned}$$

Notice that in order to measure the change in V effected by parallel translation, we made use of an orthonormal moving frame X_1, X_2 defined on all of N . Such a moving frame always exists, because a unit vector field X_1 exists on the *bounded* manifold N (see Problem I.11-13(e)), and X_2 is then determined by the orientation.

In our next theorem, the region N must be very special.

5. THEOREM. Let M be an oriented 2-dimensional Riemannian manifold, with Gaussian curvature K , and volume element dA . Let $N \subset M$ be a compact 2-dimensional manifold-with-boundary which is diffeomorphic to a subset of \mathbb{R}^2 , and whose boundary is connected. Let ds be the volume element of ∂N (determined by the induced Riemannian metric and induced orientation of ∂N), and let κ_g be the signed geodesic curvature of ∂N (on which we have a direction determined by the induced orientation). Then

$$\int_N K dA = - \int_{\partial N} \kappa_g ds + 2\pi.$$

PROOF. Because of our hypotheses on N , we might as well assume that M is an open subset of \mathbb{R}^2 . On M we define a positively oriented orthonormal moving frame X_1, X_2 by requiring X_1 to be a positive multiple of $\partial/\partial x^1$. Let $c: [a, b] \rightarrow \partial N$ be a closed curve, parameterized by arclength σ , such that $c'(\sigma)$ is positively oriented, and let ϕ be a continuous choice of the angle between X_1

and $c'(\sigma)$. Then

$$\begin{aligned}
 \int_N K \, dA &= \int_N K \theta^1 \wedge \theta^2 = - \int_N d\omega_1^2 = - \int_{\partial N} \omega_1^2 \\
 &= - \int_a^b \omega_1^2(c'(\sigma)) \, d\sigma \\
 &= - \int_a^b \kappa_g(\sigma) \, d\sigma + \int_a^b \phi'(\sigma) \, d\sigma \quad \text{by Corollary 3} \\
 &= - \int_{\partial N} \kappa_g \, ds + \phi(b) - \phi(a).
 \end{aligned}$$

To complete the proof, we have to show that $\phi(b) - \phi(a) = 2\pi$. This is done by noting three things.

- (1) The number $\phi(b) - \phi(a)$ is a multiple of 2π , since $\phi(b)$ and $\phi(a)$ are both choices of the angle for $c'(a) = c'(b)$.
- (2) Let $\langle \cdot, \cdot \rangle$ be the Riemannian metric on M , and let $\langle \cdot, \cdot \rangle$ be the usual Riemannian metric on \mathbb{R}^2 . It is easily checked that for each $t \in [0, 1]$, the tensor

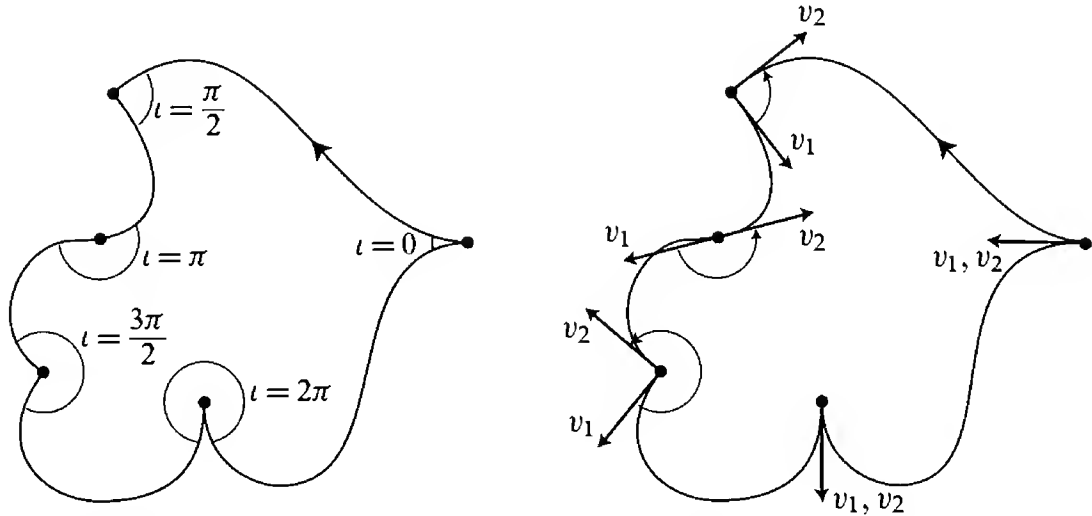
$$\langle \cdot, \cdot \rangle^t = t \langle \cdot, \cdot \rangle + (1-t) \langle \cdot, \cdot \rangle$$

is also a Riemannian metric. Let X^t_1, X^t_2 be a positively oriented moving frame which is orthonormal with respect to $\langle \cdot, \cdot \rangle^t$, and for which X^t_1 is a positive multiple of $\partial/\partial x^1$; then let ϕ^t be a continuous choice of the angle between X^t_1 and $c'(\sigma)/\|c'(\sigma)\|^t$. The choice ϕ^t clearly depends continuously on t (if we make $\phi^t(a)$ vary continuously). Consequently, $\phi^t(b) - \phi^t(a)$ varies continuously with t . Since it is always a multiple of 2π , it must be constant.

- (3) When $t = 0$, we have $X^0_1 = \partial/\partial x^1$ and $X^0_2 = \partial/\partial x^2$, and consequently ϕ^0 is just a choice of the angle between the x -axis and c' . So $\phi^0(b) - \phi^0(a)$ is the total curvature of c , as defined on pg. II.18. By the *Hopf Umlaufsatz* (Theorem II.1-7), this total curvature is 2π . ♦

We would like to generalize Theorem 5 slightly, so that the boundary of N need not be smooth, but only piecewise smooth. The proof itself will go through almost precisely as before, and the real problem is to formulate the definitions and state the results correctly (something almost no one ever bothers to do). We will say that a compact 2-dimensional manifold-with-boundary $N \subset M$ is a **polygon** if ∂N is connected and if there is a simple closed curve $c: [a, b] \rightarrow N$ such that c is a smooth imbedding on each interval $[t_{i-1}, t_i]$ of some partition $a = t_0 < \dots < t_{n+1} = b$. Thus $c'(t)$ exists for all $t \neq t_i$, and the right and left hand derivatives $c'(t_i^+)$ and $c'(t_i^-)$ exist for all t_i . It will be convenient to work only with curves c such that $c'(b^-) = c'(a^+)$. The **vertices** of c will then

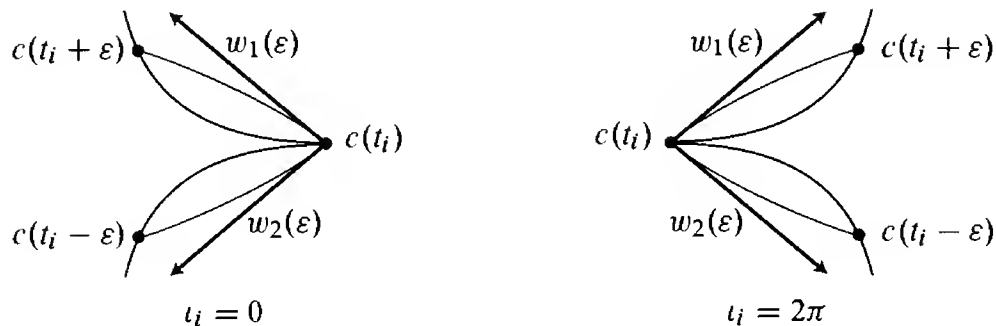
be t_1, \dots, t_n ; for each such vertex t_i , we would like to define its **interior angle** $\iota_i \in [0, 2\pi]$, as shown below. To do this, we first choose c so that $c'(t)$ is always



positively oriented. Let $v_1 = c'(t_i^+)$ and $v_2 = -c'(t_i^-)$. If v_1 and v_2 do not point in the same direction, we define

$\iota_i =$ oriented angle (between 0 and 2π) from v_1 to v_2 .

This still leaves us with the problem of defining ι_i when v_1 and v_2 point in the same direction. To treat this case, let $w_1(\varepsilon)$ be the tangent vector of the geodesic from $c(t_i)$ to $c(t_i + \varepsilon)$, and let $w_2(\varepsilon)$ be the tangent vector of the geodesic from $c(t_i)$ to $c(t_i - \varepsilon)$. For sufficiently small $\varepsilon > 0$, the vectors $w_1(\varepsilon)$

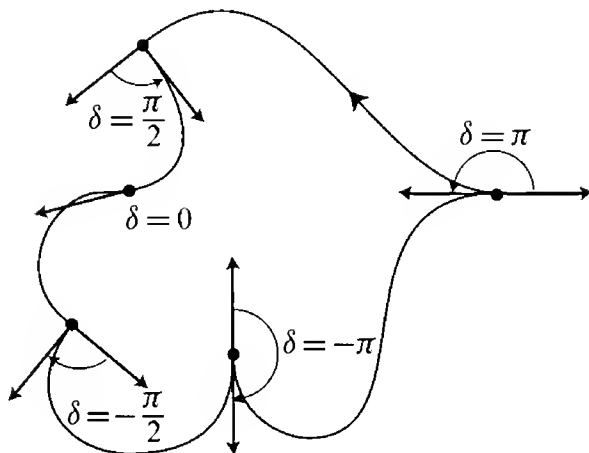


and $w_2(\varepsilon)$ are nearly in the same direction, so, in particular, they do not point in opposite directions. Therefore, the orientation of $(w_1(\varepsilon), w_2(\varepsilon))$ cannot change, since $w_1(\varepsilon)$ and $w_2(\varepsilon)$ are always distinct for small ε . We define ι_i to be 0 if $(w_1(\varepsilon), w_2(\varepsilon))$ is positively oriented, and 2π if it is negatively oriented. [The formula

$$\iota_i = \lim_{\varepsilon \rightarrow 0} \{\text{oriented angle from } w_1(\varepsilon) \text{ to } w_2(\varepsilon)\}$$

could be used as a general definition that would work in all cases.] Now that we have successfully defined the interior angle ι_i , we define the **discontinuity** δ_i of c' at t_i by

$$\delta_i = \pi - \iota_i \in [-\pi, \pi].$$



Remark: It is easy to see that if ϕ is an angle between $c'(t_i^-)$ and some vector $X \in M_{c(t_i)}$, then $\phi + \delta_i$ is an angle between $c'(t_i^+)$ and X .

6. THEOREM (THE GAUSS-BONNET FORMULA). Let M be an oriented 2-dimensional Riemannian manifold, with Gaussian curvature K , and volume element dA . Let $N \subset M$ be a polygon which is diffeomorphic to a subset of \mathbb{R}^2 , let ds be the volume element of ∂N , and let κ_g be its signed geodesic curvature. Suppose that ∂N has vertices t_1, \dots, t_n , with discontinuities $\delta_1, \dots, \delta_n$. Then

$$\begin{aligned} \int_N K dA &= - \int_{\partial N} \kappa_g ds - \sum_{i=1}^n \delta_i + 2\pi \\ &= - \int_{\partial N} \kappa_g ds + \sum_{i=1}^n \iota_i + (2 - n)\pi. \end{aligned}$$

PROOF. As in the proof of the previous theorem, we assume that M is an open subset of \mathbb{R}^2 , and we define X_1, X_2 exactly as before. Choose the curve $c: [a, b] \rightarrow \partial N$ to be parameterized by arclength σ , and so that $c'(\sigma)$ is positively oriented. By our Remark, we can choose $\phi_i: [t_{i-1}, t_i] \rightarrow \mathbb{R}$ so that each ϕ_i is a continuous choice of the angle between X_1 and $c'(\sigma)$ on (t_{i-1}, t_i) , and so that

$$(*) \quad \phi_{i+1}(t_i) - \phi_i(t_i) = \delta_i \quad i = 1, \dots, n.$$

Then

$$\begin{aligned}
 \int_N K dA &= \int_N K d\theta^1 \wedge d\theta^2 = - \int_N d\omega_1^2 = - \int_{\partial N} \omega_1^2 \\
 &= - \sum_{i=1}^{n+1} \int_{t_{i-1}}^{t_i} \omega_1^2(c'(\sigma)) d\sigma \\
 &= - \sum_{i=1}^{n+1} \left[\int_{t_{i-1}}^{t_i} \kappa_g(\sigma) d\sigma - \int_{t_{i-1}}^{t_i} \phi_i'(\sigma) d\sigma \right] \quad \text{by Corollary 3} \\
 &= - \int_{\partial N} \kappa_g ds + \sum_{i=1}^{n+1} \phi_i(t_i) - \phi_i(t_{i-1}) \\
 &= - \int_{\partial N} \kappa_g ds - \sum_{i=1}^n \delta_i + [\phi_{n+1}(b) - \phi_1(a)], \quad \text{using (*).}
 \end{aligned}$$

To complete the proof, we have to show that $\phi_{n+1}(b) - \phi_1(a) = 2\pi$. We do this by showing that the three observations in the previous proof now hold for $\phi_{n+1}(b) - \phi_1(a)$.

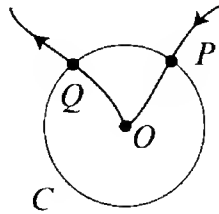
(1) and (2) are obvious.

(3) We are now dealing with a piecewise smooth simple closed curve c in \mathbb{R}^2 . We want to show that

$$2\pi = \phi_{n+1}(b) - \phi_1(a) = \sum_{i=1}^{n+1} \phi_i(t_i) - \phi_i(t_{i-1}) + \sum_{i=1}^n \delta_i.$$

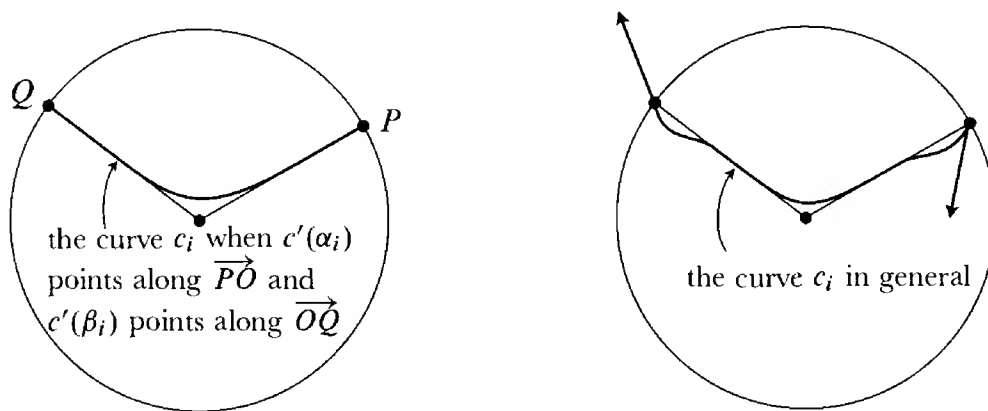
To prove this generalization of the Hopf Umlaufsatz, we proceed as follows.

Draw a small circle C around $O = c(t_i)$. Let $P = c(\alpha_i)$ be the last point of $c|_{[t_{i-1}, t_i]}$ on C and let $Q = c(\beta_i)$ be the first point of $c|_{[t_i, t_{i+1}]}$ on C . The



oriented angle from \overrightarrow{OQ} to \overrightarrow{OP} approaches ι_i as the radius of C approaches 0. Also, the direction of $c'(\alpha_i)$ is nearly that of \overrightarrow{PO} , while the direction of $c'(\beta_i)$

is nearly that of \overrightarrow{OQ} . The picture below shows a curve c_i from P to Q which begins in the direction of $c'(\alpha_i)$, ends in the direction of $c'(\beta_i)$, and stays inside C except at P and Q . The total change in angle of the tangent vector c_i'



is easily seen to be very close to $\delta_i = \pi - \iota_i$. For each i , let us replace the portion of c between α_i and β_i by the curve c_i . If the circles C are small enough (so that curve does not enter the circle around $c(t_i)$ on any interval other than $[t_{i-1}, t_{i+1}]$), then the new curve, \tilde{c} , is simple. The total change in angle of the tangent vector \tilde{c}' is therefore 2π , by the Hopf Umlaufsatz. On the other hand, the total change along the portions c_i adds up to something very close to $\sum_i \delta_i$, while the total change along the other portions adds up to something very close to $\sum_i \phi_i(t_i) - \phi_i(t_{i-1})$. Therefore the number

$$\sum_{i=1}^{n+1} \phi_i(t_i) - \phi_i(t_{i-1}) + \sum_{i=1}^n \delta_i = \phi_{n+1}(b) - \phi_1(a)$$

must be close to 2π . Therefore it must be exactly 2π . ♦

7. COROLLARY. If the sides of the polygon N in Theorem 6 are geodesics, then

$$\int_N K dA = - \sum_{i=1}^n \delta_i + 2\pi = \sum_{i=1}^n \iota_i + (2 - n)\pi.$$

In particular, for a geodesic triangle we have

$$\int_N K dA = \iota_1 + \iota_2 + \iota_3 - \pi.$$

We are now in a position to find the integral of $K dA$ over all of M . The first method of doing this will use a triangulation $\{\sigma_i\}$ of M by 2-simplexes σ_i . Triangulations are defined in the “optional” Chapter 11 of Volume I (see pg. I.426),

and it is not easy to prove that they always exist on C^∞ manifolds; however, the proof for compact 2-manifolds is fairly easy (Problem 4-17). For a given triangulation of M we will let

V = number of 0-simplexes ("vertices")

E = number of 1-simplexes ("edges")

F = number of 2-simplexes ("faces").

The number

$$V - E + F = \chi(M)$$

is called the **Euler characteristic** of M . According to Theorem I.11-5, we have

$$\chi(M) = \dim H^0(M) - \dim H^1(M) + \dim H^2(M),$$

so actually $\chi(M)$ does not depend on the triangulation. However, it is not necessary to know this fact in order to follow the next proof.

8. THE GAUSS-BONNET THEOREM. Let M be a compact oriented 2-dimensional Riemannian manifold, with Gaussian curvature K , and volume element dA . Then

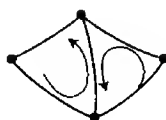
$$\int_M K dA = 2\pi \cdot \chi(M).$$

PROOF. Consider a triangulation $\sigma_1, \dots, \sigma_F$ of M . Let A_j, B_j, C_j be the three interior angles of σ_j . Then Theorem 6 gives

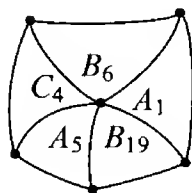
$$\begin{aligned} \int_M K dA &= \sum_{j=1}^F \int_{\sigma_j} K dA \\ &= \sum_{j=1}^F \left(\int_{\partial\sigma_j} \kappa_g ds \right) + \sum_{j=1}^F (A_j + B_j + C_j) - \sum_{j=1}^F 3\pi + \sum_{j=1}^F 2\pi. \end{aligned}$$

Now we note the following.

- (1) The sum $\sum \int \kappa_g ds$ is 0, because each edge of the triangulation appears twice, with opposite orientations.



- (2) The sum $\sum(A_j + B_j + C_j)$ is $2\pi V$, since the sum of all interior angles occurring at each vertex is exactly 2π .



- (3) The sum $-\sum 3\pi$ is just $-3F\pi$. On the other hand, we clearly have $3F = 2E$, since $3F$ is the total number of edges of all faces, each edge being counted twice since it is in two faces. So $-\sum 3\pi = 2\pi(-E)$.

- (4) The sum $\sum 2\pi$ is just $2\pi F$.

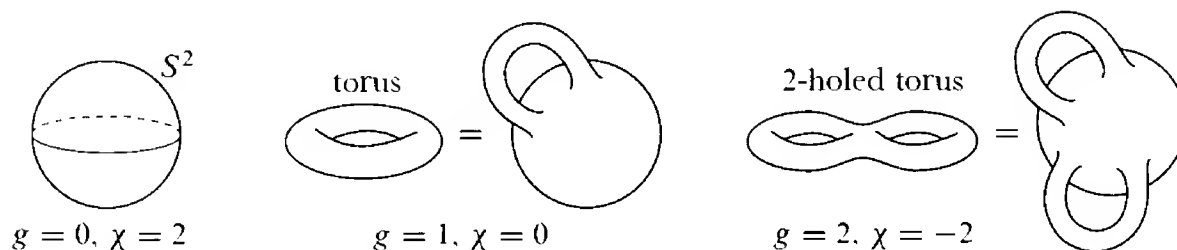
Therefore, our final sum is $2\pi(V - E + F) = 2\pi \cdot \chi(M)$. ♦

A whole slew of consequences follows immediately from this spectacular theorem, which expresses a differential-geometric quantity $\int_M K dA$ in terms of a number which has nothing at all to do with curvature, or even with a Riemannian metric. Note, first of all, that we can restate our result in a way which does not involve $\chi(M) = \dim H^0(M) - \dim H^1(M) + \dim H^2(M)$. We have shown that for *any* triangulation of M we have

$$\int_M K dA = 2\pi(V - E + F).$$

Picking a fixed triangulation, this shows that $\int_M K dA$ is independent of the metric. On the other hand, picking a fixed metric, we obtain an “elementary” proof, without using cohomology, that $V - E + F$ is independent of the triangulation.

The simplest consequences of the Gauss-Bonnet Theorem involve the sign of K on various compact oriented surfaces. All such surfaces are homeomorphic to the surface obtained by adding $g \geq 0$ handles to S^2 , and the Euler characteristic is then given by $\chi = 2 - 2g$. The latter result may be proved as



in Problem I.11-2, or by considering triangulations (we just have to check that $V - E + F$ changes by -2 when we add one more handle). Therefore,

$$\begin{aligned} \int_M K dA &> 0 \quad \text{if } M \text{ is homeomorphic to } S^2 \\ \int_M K dA &= 0 \quad \text{if } M \text{ is homeomorphic to a torus} \\ \int_M K dA &< 0 \quad \text{if } M \text{ is any other compact oriented surface.} \end{aligned}$$

It follows, in particular, that if there is a metric on a compact oriented 2-manifold M with $K > 0$ everywhere, then M must be homeomorphic to S^2 . This result is rather different from Theorem 2-11, because we do not assume that the metric comes from an imbedding in \mathbb{R}^3 (whether such a metric does, in fact, always come from an imbedding is a question which we will mention only later, in Chapter 11). On the other hand, if there is a metric on M with $K = 0$ everywhere, then M must be homeomorphic to the torus. As we have already seen (pg. II.179), such a flat metric does indeed exist on the torus. It is a good exercise to compute that $\int_M K dA = 0$ when M is the torus on page 159. Finally, if there is a metric on M with $K < 0$ everywhere, then M must be a sphere with $g \geq 2$ handles. Of course, we can never find an imbedding of such a surface M into \mathbb{R}^3 with $K < 0$ everywhere (by Proposition 2-8). But there is, nevertheless, an abstract Riemannian metric on M which has $K < 0$ everywhere; in fact, there is a Riemannian metric on M with $K = -1$ everywhere. The construction of such metrics is carried out in Addendum 1.

Our next consequence of the Gauss-Bonnet Theorem involves the notion of the index of a vector field, as defined in Chapter I.11.

9. THEOREM. Let M be a compact oriented 2-dimensional Riemannian manifold with Gaussian curvature K , and volume element dA . Let X be a vector field on M with only finitely many zeros. Then

$$\int_M K dA = 2\pi \cdot (\text{sum of indices of } X).$$

FIRST PROOF. We just combine the Gauss-Bonnet Theorem,

$$\int_M K dA = 2\pi \cdot \chi(M),$$

with the Poincaré-Hopf Theorem (I.11-30),

$$\chi(M) = \text{sum of indices of } X.$$

SECOND PROOF. This proof does not involve the intermediary $\chi(M)$, which appears nowhere in the statement of the theorem, nor will a triangulation be invoked. Let p_1, \dots, p_r be the zeros of X , and choose disjoint closed discs D_i containing p_i . Each D_i is diffeomorphic to $D = \{x \in \mathbb{R}^2 : |x| \leq 1\}$, and we will let $D_i(\varepsilon)$ denote the set corresponding to $\{x \in \mathbb{R}^2 : |x| \leq \varepsilon\}$. Let

$$N(\varepsilon) = M - (\bigcup_i \text{interior } D_i(\varepsilon)).$$

On $N(\varepsilon)$ there is a positively oriented orthonormal moving frame X_1, X_2 with $X_1 = X/\|X\|$. Then

$$\begin{aligned} \int_{N(\varepsilon)} K dA &= - \int_{N(\varepsilon)} d\omega_1^2 \\ &= - \sum_i \int_{\partial D_i(\varepsilon)} \omega_1^2 \quad \text{by Stokes' Theorem.} \end{aligned}$$

For the moment consider one particular i . Let X'_1, X'_2 be a fixed positively oriented orthonormal moving frame on D_i . On D_i minus a line segment we have a differentiable choice θ of the angle between X_1 and X'_1 , and by Proposition 1 we have

$$\begin{aligned} - \int_{\partial D_i(\varepsilon)} \omega_1^2 &= \int_{\partial D_i(\varepsilon)} d\theta - \int_{\partial D_i(\varepsilon)} \omega'^2_1 \\ &= (\text{index of } X \text{ at } p_i) - \int_{\partial D_i(\varepsilon)} \omega'^2_1 \\ &\quad [\text{where } \omega'^2_1 \text{ really depends on } i]. \end{aligned}$$

Therefore,

$$\lim_{\varepsilon \rightarrow 0} - \int_{\partial D_i(\varepsilon)} \omega_1^2 = (\text{index of } X \text{ at } p_i) - 0.$$

Consequently,

$$\int_M K dA = \lim_{\varepsilon \rightarrow 0} \int_{N(\varepsilon)} K dA = \sum_i \text{index of } X \text{ at } p_i. \spadesuit$$

Notice that the second proof of Theorem 9 reproves the fact that the sum of the indices of a vector field X is independent of X (it also reproves the fact that $\int_M K dA$ is independent of the metric). In conjunction with the proof of the Gauss-Bonnet Theorem, it reproves the fact that the sum of the indices of a vector field is $V - E + F$.

In contrast to the previous consequence of the Gauss-Bonnet Theorem, in which $\chi(M)$ does not appear, in the next consequence K does not appear.

10. THEOREM. Let $M \subset \mathbb{R}^3$ be a compact oriented 2-dimensional manifold, and let $\nu: M \rightarrow S^2$ be the normal map. Then

$$\text{degree of } \nu = \frac{\chi(M)}{2}.$$

PROOF. By the definition of $\deg \nu$ (pg. I.275), we have

$$\int_M \nu^* \omega = \deg \nu \cdot \int_{S^2} \omega$$

for all 2-forms ω on S^2 . Choosing ω to be the volume element da of S^2 , this means that

$$\begin{aligned} (4\pi) \deg \nu &= \int_M \nu^*(da) \\ &= \int_M K dA, \end{aligned}$$

where K is the curvature for the induced metric. Together with the Gauss-Bonnet Theorem, this yields the desired result. ♦

Just as in the case of Theorem 9, one would expect to find a proof of Theorem 10 which does not use the intermediary K . In fact, the same statement can be proved for hypersurfaces of \mathbb{R}^n , and this result played an important role in generalizing the Gauss-Bonnet Theorem to higher dimensions (see Chapter 13). Because the proof is differential-topological in nature, rather than differential-geometric, it has been shunted off to Addendum 2.

Our next result is merely a curiosity, an alternative proof of Theorem 2-11 which avoids covering spaces.

11. **PROPOSITION.** Let M be a compact connected 2-manifold, and let $f: M \rightarrow \mathbb{R}^3$ be an immersion with $K(p) > 0$ for all $p \in M$. Then M is orientable, the normal map $N: M \rightarrow S^2 \subset \mathbb{R}^3$ is a diffeomorphism, the map $f: M \rightarrow \mathbb{R}^3$ is an imbedding, and $f(M)$ is convex.

PROOF. Orientability of M is trivial, as in the proof of Theorem 2-11. The Gauss-Bonnet Theorem then shows that

$$2\pi \cdot \chi(M) = \int_M K \, dA > 0.$$

The only possibility is that M is homeomorphic to S^2 , with $\chi(M) = 2$; so

$$\int_M K \, dA = 4\pi.$$

Since $N(M) \subset S^2$ is closed (by compactness of M) and also open (as $K(p) \neq 0$ means that N is regular at p), the map N is onto S^2 . To prove that N is one-one, suppose instead that $N(p) = N(q)$ for some $p \neq q \in M$. Then there is an open set $U \ni q$ such that $N(M - \bar{U}) = S^2$. If da is the volume element on S^2 , then for any open set $V \subset M - \bar{U}$ on which N is one-one we have

$$\int_V K \, dA = \int_V N^*(da) = \int_{N(V)} da, \quad \text{since } N \text{ is orientation preserving.}$$

It follows that

$$\int_{M - \bar{U}} K \, dA \geq \int_{S^2} da = 4\pi.$$

Therefore

$$\int_M K \, dA = \int_{M - \bar{U}} K \, dA + \int_U K \, dA > 4\pi,$$

a contradiction. So N is one-one. The remainder of the proof is the same as for Theorem 2-11. ♦

Finally, here's a result that isn't about surfaces at all! (For a history of this result, see McCleary [1].)

12. **THEOREM (JACOBI; 1842).** Let c be a closed curve in \mathbb{R}^3 with nowhere vanishing curvature κ , and let \mathbf{n} be its normal map, into S^2 . If \mathbf{n} is a simple closed curve on S^2 , then it divides S^2 into two regions of equal area.

PROOF. Let B be one of the regions into which \mathbf{n} divides S^2 , and orient $c: [a, b] \rightarrow S^2$ so that the corresponding orientation for \mathbf{n} coincides with the induced orientation for ∂B . Theorem 5 gives

$$\int_B dA = - \int_{\partial B} \kappa_g d\sigma + 2\pi,$$

where κ_g is the geodesic curvature of \mathbf{n} on S^2 , and σ is the arclength function of \mathbf{n} . Since S^2 has area 4π , it suffices to prove that the integral of κ_g is 0. Now for any arclength parameterized curve γ on a surface M , the definition of its geodesic curvature κ_g is

$$\mathbf{T}\gamma'' = \kappa_g \cdot \nu \times \gamma',$$

where ν is the normal to M . This implies that

$$\kappa_g = \langle \gamma'', \nu \times \gamma' \rangle = \langle \gamma' \times \gamma'', \nu \rangle.$$

When γ is not parameterized by arclength, we have

$$\kappa_g(t) = \langle \gamma'(t) \times \gamma''(t), \nu(\gamma(t)) \rangle / |\gamma'(t)|^3.$$

Applying this to our curve \mathbf{n} on S^2 we obtain

$$\kappa_g(s) = \langle \mathbf{n}'(s) \times \mathbf{n}''(s), \mathbf{n}(s) \rangle / \left(\frac{d\sigma}{ds} \right)^3.$$

The Serret-Frenet formulas for c allow us to write this as

$$\kappa_g(s) = [\kappa(s)\tau'(s) - \kappa'(s)\tau(s)] / \left(\frac{d\sigma}{ds} \right)^3,$$

where κ and τ are the curvature and torsion of c . Now since

$$\mathbf{n}' = -\kappa\mathbf{t} + \tau\mathbf{b},$$

we have

$$\frac{d\sigma}{ds} = |\mathbf{n}'(s)| = \sqrt{\kappa^2(s) + \tau^2(s)}.$$

So

$$\begin{aligned} \kappa_g(s) &= \frac{\kappa\tau' - \kappa'\tau}{\kappa^2 + \tau^2} / \frac{d\sigma}{ds} = \frac{d}{ds} \left(\arctan \frac{\tau}{\kappa} \right) \frac{d\sigma}{ds} \\ &= \frac{d}{d\sigma} \left(\arctan \frac{\tau}{\kappa} \right). \end{aligned}$$

Since c is a closed curve, this gives

$$\int_{\partial B} \kappa_g d\sigma = \int_a^b d \left(\arctan \frac{\tau}{\kappa} \right) = 0. \quad \blacklozenge$$

We will conclude this Chapter with some considerations which, although they do not bear directly on the Gauss-Bonnet Theorem, are nevertheless related to the integral of the curvature. As we have already noted in the proof of Theorem 10, for an immersion $i: M \rightarrow \mathbb{R}^3$ of a compact oriented 2-manifold M into \mathbb{R}^3 , the integral of its curvature K is

$$\int_M K dA = (4\pi) \deg N,$$

where $N: M \rightarrow S^2$ is the normal map. This degree of N is simply the “signed” number of points in $N^{-1}(p)$ for any regular value $p \in S^2$ of N (Theorem I.8-12). Now let us consider the actual number $\#(p) \geq 0$ of points in $N^{-1}(p)$. We would like to look at

$$\int_{S^2} \# \cdot da,$$

where da is the volume element of S^2 ; for technical reasons we will instead consider

$$\int_{S^2-C} \# \cdot da,$$

where C is the set of critical values of N (which has measure 0 by Sard’s Theorem). Notice that for some $p \in S^2$ the set $N^{-1}(p)$ might be infinite; but such p are clearly contained in C and therefore do not bother us. Moreover, on $S^2 - C$ the function $\#$ is locally constant, so it is certainly continuous, and therefore the integral makes sense. Any point of $S^2 - C$ has a neighborhood U on which $\#$ has a constant value m and such that $N^{-1}(U) \subset M$ is the disjoint union of m open sets V_1, \dots, V_m on each of which $N: V_\alpha \rightarrow U$ is a diffeomorphism. Since $N^*(da) = K dA$, we have

$$\int_{V_\alpha} K dA = \begin{cases} \int_U da & N \text{ orientation preserving, i.e., } K > 0 \\ - \int_U da & N \text{ orientation reversing, i.e., } K < 0. \end{cases}$$

So

$$\int_{V_\alpha} |K| dA = \int_U da.$$

From this we readily see that

$$\int_M |K| dA = \int_{S^2-C} \# da.$$

This number is called the **total absolute curvature** of the immersion $i: M \rightarrow \mathbb{R}^3$.

13. THEOREM. For any immersion $i: M \rightarrow \mathbb{R}^3$ of a compact oriented 2-manifold in \mathbb{R}^3 , the total absolute curvature is $\geq 4\pi$. In fact,

$$\int_{\{p \in M: K(p) > 0\}} |K| dA = \int_{\{p \in M: K(p) > 0\}} K dA \geq 4\pi.$$

If the total absolute curvature of the immersion $i: M \rightarrow \mathbb{R}^3$ equals 4π , then i is an imbedding and $i(M)$ is convex.

PROOF. Let $M' = \{p \in M : K(p) \geq 0\}$. Since

$$\int_{\{p \in M: K(p) > 0\}} K dA = \int_{M'} K dA = \int_{M'} N^*(da),$$

the first part of the theorem will certainly be proved if we show that $N(M') = S^2$. For any $v \in S^2$, consider a plane $P \subset \mathbb{R}^3$ perpendicular to v and far away from $i(M)$ in the direction of v . Move this plane towards 0 until it first touches $i(M)$ at a point p . Then $N(p) = v$. Moreover, $K(p) \geq 0$, since M does not lie on both sides of its tangent plane P_0 at p . This proves the first part of the Theorem.

Now suppose that the total absolute curvature of the immersion $i: M \rightarrow \mathbb{R}^3$ equals 4π . We will first show that $i(M)$ lies on one side of each of its tangent planes. Suppose that M_p cuts $i(M)$, i.e., that $i(M)$ lies on both sides of its tangent plane M_p . The points furthest from M_p on each side will have normals which are negatives of each other and both perpendicular to M_p . Since $N(p)$ is also perpendicular to M_p , the point $v = N(p)$ has $\#(v) \geq 2$. It is clear that $i(M)$ also lies on both sides of the tangent planes M_q for q in a neighborhood U of p . Now if $K(p) \neq 0$, then we can also assume that n is a diffeomorphism on U , by making U smaller if necessary. Then $\# \geq 2$ on the whole open set $N(U)$. Since we have already shown that $\# \geq 1$ on S^2 , this shows that

$$\int_M |K| dA = \int_{S^2} \# da > 4\pi,$$

a contradiction. So to complete the proof that $i(M)$ lies on one side of M_p , we just have to show that if $K(p) = 0$, then there would be some other point \bar{p} such that $i(M)$ lies on both sides of $M_{\bar{p}}$, and also $K(\bar{p}) \neq 0$.

We will first find a point p' whose tangent plane cuts $i(M)$ and such that p' is not a planar point (i.e., either $K(p') \neq 0$ or p' is a parabolic point). There is certainly no problem doing this if there are non-planar points arbitrarily close

to p . So suppose that all points in a neighborhood of p are planar points; the image of this neighborhood under the immersion i then lies in M_p . Consider the set of all points in M whose image lies in M_p , and the component of this set which contains p . Let q be a boundary point of this component. Then $M_q = M_p$, so M_q also cuts $i(M)$. Moreover, q cannot have a whole neighborhood of planar points (for then it would not be a boundary point of the component). So there are points p' arbitrarily close to q such that p' is not a planar point. By choosing p' close enough, we can insure that $M_{p'}$ cuts $i(M)$.

If the point p' which we have produced satisfies $K(p') \neq 0$, we are done. Suppose instead that p' is a parabolic point. Let $L_{p'}$ be the straight line given by Corollary 5-6. Consider the set of points on $L_{p'}$ where $K = 0$; it is a certain closed interval I (which might be just $\{p'\}$). Along I the tangent plane of $i(M)$ is constant, by Proposition 4-5. So if q' is an endpoint of I , then $M_{q'} = M_{p'}$ cuts $i(M)$. Now q' cannot have a whole neighborhood of parabolic points, so there are points \bar{p} arbitrarily close to q' with $K(\bar{p}) \neq 0$. By choosing \bar{p} close enough, we can insure that $M_{\bar{p}}$ cuts $i(M)$.

Thus we have shown that $i(M)$ lies on one side of each of its tangent planes. Let $C \subset \mathbb{R}^3$ be the intersection of the closed half-spaces which are bounded by the planes M_p and contain the points of $i(M)$. Then C is a compact convex set with non-empty interior, and $i(M) \subset \text{boundary } C$. Since $i: M \rightarrow \mathbb{R}^3$ is an immersion, the set $i(M)$ is both open and closed in boundary C , hence $i(M) = \text{boundary } C$. By Problem 2-4, the map $i: M \rightarrow \text{boundary } C \approx S^2$ is a covering map. So i must be a homeomorphism. ♦

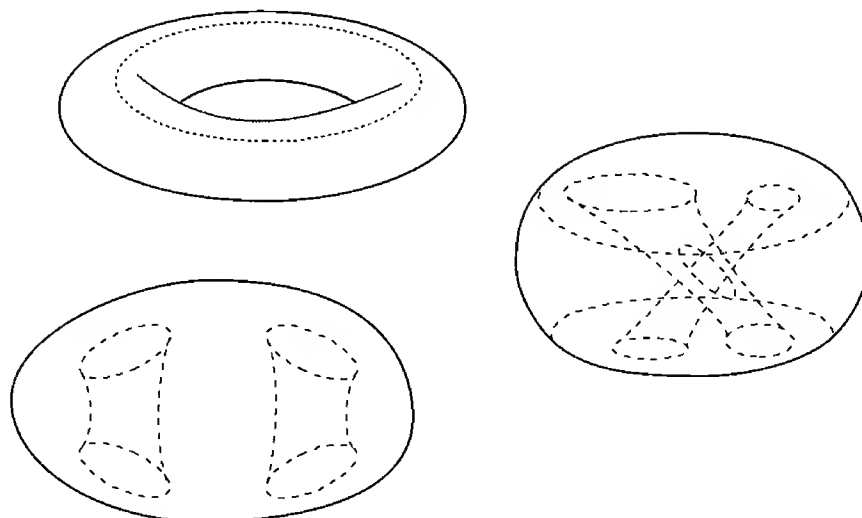
The first part of the preceding proof actually shows more than the asserted inequality, for we clearly have strict inequality if any open subset of S^2 is covered twice. Thus,

$$(*) \quad \int_{\{p \in M: K(p) > 0\}} K dA = 4\pi \iff N \text{ is one-one on } \{p \in M: K(p) > 0\}.$$

[Condition $(*)$ can also be expressed in terms of the total absolute curvature, for

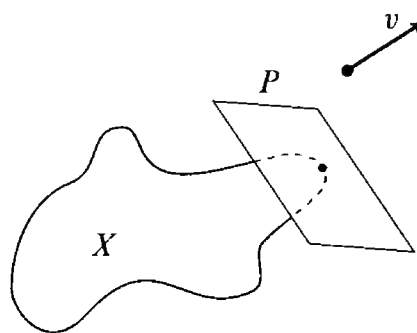
$$\begin{aligned} \int_M |K| dA &= \int_{\{p \in M: K(p) > 0\}} K dA - \int_{\{p \in M: K(p) < 0\}} K dA \\ &= 2 \int_{\{p \in M: K(p) > 0\}} K dA - \int_M K dA \\ &\geq 8\pi - 2\pi \cdot \chi(M) \quad \text{by the Gauss-Bonnet Theorem,} \end{aligned}$$

with equality if and only if $(*)$ holds.] The pictures below illustrate some immersed surfaces with property $(*)$. In each case, the region of positive curvature



is a subset of a convex surface, bounded by convex plane curves. We will show, by elementary but involved arguments, that this is always so.

Given any set $X \subset \mathbb{R}^3$, let $H(X)$ be its convex hull, the smallest convex set containing X . Elementary considerations (Problem 1) show that if X is compact, then so is $H(X)$. For any unit vector $v \in S^2 \subset \mathbb{R}^3$, consider the subset of X where the function $\iota_v(x) = \langle x, v \rangle$ has its maximum value on X . This will be a subset of the plane P perpendicular to v which is furthest from the origin and still hits X . It is called the “topset” of X in the direction v , and is clearly a subset of the boundary $\partial H(X)$ of $H(X)$.



14. LEMMA. Let $i: M \rightarrow \mathbb{R}^3$ be an immersion of a compact oriented surface satisfying $(*)$. If $i(p) \in \partial H(i(M))$, then $K(p) \geq 0$; and conversely if we have the strict inequality $K(p) > 0$, then $i(p) \in \partial H(i(M))$.

PROOF. The first assertion is clear, since $i(M)$ has no support plane at $i(p)$ if $K(p) < 0$, while $H(i(M))$ has a support plane at every point of its boundary.

For each $v \in S^2$, consider the topset of $i(M)$ in the direction v . It is a closed convex subset of a plane. Suppose it contains at least 2 points $i(p_1), i(p_2)$ for $p_1, p_2 \in M$. We clearly have $K(p_1), K(p_2) \geq 0$. Condition (*) then shows that we must have either $K(p_1) = 0$ or $K(p_2) = 0$. Thus $v \in N(\{p \in M : K(p) = 0\}) = C$, say. In other words, if $v \in S^2 - C$, then the topset in the direction v contains only one point. This point of $\partial H(i(M))$ must be $i(p)$ for some $p \in M$, and clearly $N(p) = v$ and $K(p) \geq 0$. Thus

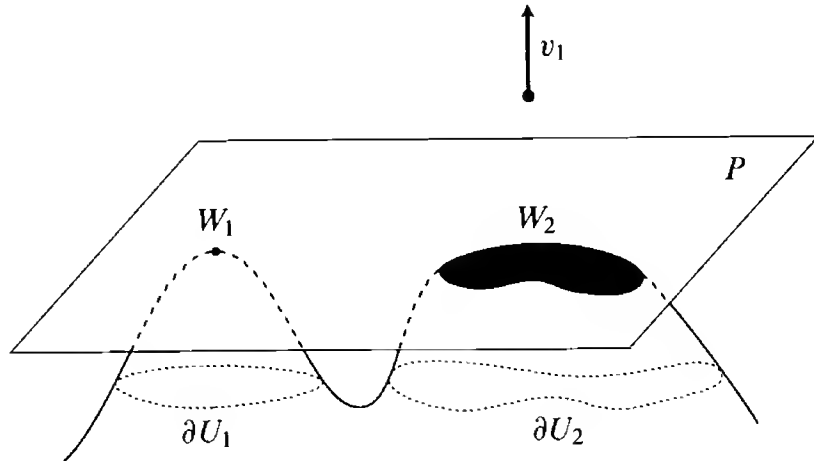
$$N(\{p \in M : i(p) \in \partial H(i(M)) \text{ and } K(p) \geq 0\}) \supset S^2 - C,$$

which has area 4π . So we cannot have $K(p) > 0$ for any other points $p \in M$. ♦

The main part of the argument goes into the following

15. LEMMA. Let $i: M \rightarrow \mathbb{R}^3$ be an immersion of a compact oriented surface satisfying (*). Then for any $v_1 \in S^2$, the topset of $i(M)$ in the direction v_1 is connected. Moreover, if $v_2 \in S^2$ is perpendicular to v_1 , then the topset in the direction v_2 of $\{\text{the topset of } i(M) \text{ in the direction } v_1\}$ is connected.

PROOF. If the topset in the direction v_1 is disconnected, then it is the disjoint union of two closed sets W_1, W_2 , which are also closed as subsets of $i(M)$, since the topset is a closed set. Let U_1, U_2 be disjoint open neighborhoods of W_1 and W_2 in $i(M)$. On the compact set ∂U_i we have $\langle x, v_1 \rangle$ strictly smaller than

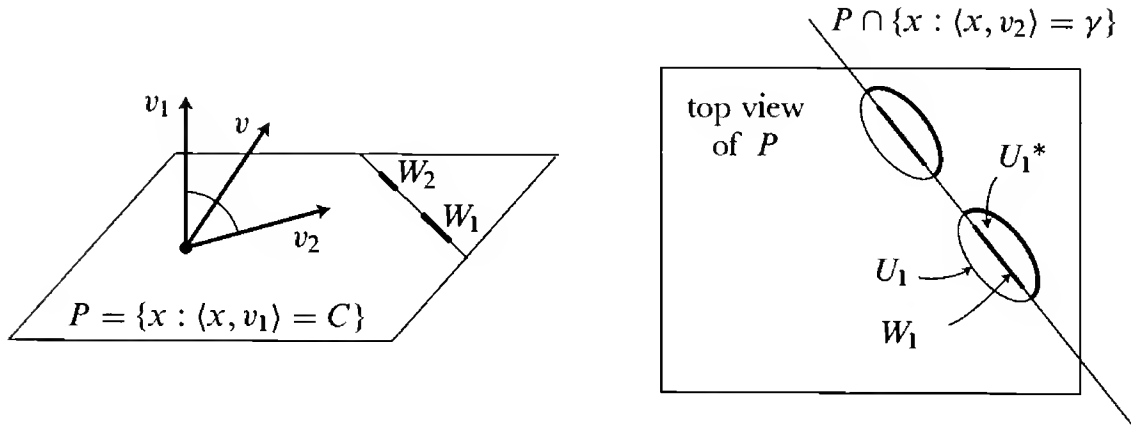


the value of $\langle x, v_1 \rangle$ for $x \in W_i$. So the same is true for all $v \in S^2$ sufficiently

close to v_1 . This means that for all such v , the topset of U_i in the direction v contains a point $i(p) \in U_i$ not on the boundary of U_i . This implies that $v = N(p)$, and also that $K(p) \geq 0$. So for all v in a whole neighborhood of v_1 , with the exception of a set of measure 0, we have $v = N(p)$ where $i(p) \in U_i$ and $K(p) > 0$. Since the U_i are disjoint, this contradicts condition (*).

Now consider the topset in the direction v_2 of the topset of $i(M)$ in the direction v_1 ; say that ι_{v_2} has the value γ on this topset. Suppose this topset is disconnected and write it as the disjoint union of closed sets W_1 and W_2 . Choose U_1, U_2 to be disjoint open neighborhoods of W_1, W_2 in $i(M)$, and let

$$U_i^* = \{x \in U_i : \langle x, v_2 \rangle \geq \gamma\}.$$



One part of ∂U_i^* is the set

$$A_i = \partial U_i \cap \{x : \langle x, v_2 \rangle \geq \gamma\}$$

(indicated by a heavy line in the figure). On this compact set we have $\langle x, v_1 \rangle$ strictly smaller than the value $\langle x, v_1 \rangle$ for $x \in W_i$. So the same is true for all unit v close to v_1 . The other part of ∂U_i^* is

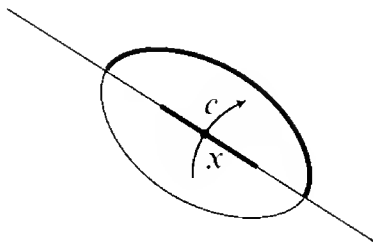
$$B_i = U_i \cap \{x : \langle x, v_2 \rangle = \gamma\}.$$

On this set, the function $\langle x, v_1 \rangle$ takes on its maximum at some $x \in W_i$. Suppose we choose our unit vector v in the plane spanned by v_1 and v_2 . Then for $x \in B_i$ we have

$$\langle x, v \rangle = (\text{constant}) \cdot \langle x, v_1 \rangle + \text{constant},$$

so the maximum still occurs on W_i . Thus, for a unit vector v_0 close enough to v_1 , and in the plane of v_1 and v_2 , the maximum of $\langle x, v_0 \rangle$ on ∂U_i^* occurs

at a point $x \in W_i$. Now we can choose a curve c in $i(M)$ with $c(0) = x$ and $c'(0) = v_2$, since the tangent plane at x is clearly the plane P which contains



the topset in the direction v_1 . If

$$v_0 = av_1 + bv_2, \quad a, b > 0,$$

then

$$\begin{aligned} \left. \frac{d}{dt} \right|_{t=0} \langle c(t), v_0 \rangle &= \langle c'(0), v_0 \rangle \\ &= \langle v_2, v_0 \rangle = b > 0. \end{aligned}$$

So $\langle c(t), v_0 \rangle$ has values greater than $\langle c(0), v_0 \rangle = \langle x, v_0 \rangle$ for small $t > 0$. This shows that the maximum of $\langle x, v_0 \rangle$ on U_i^* occurs at a point not on the boundary. Hence the same is true of $\langle x, v \rangle$ for all v sufficiently close to v_0 . As before, this leads to a contradiction. ♦

16. THEOREM. Let $i: M \rightarrow \mathbb{R}^3$ be an immersion of a compact oriented surface such that

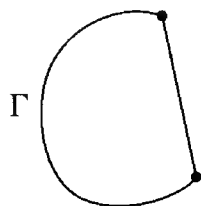
$$\int_{\{p \in M: K(p) > 0\}} K \, dA = 4\pi.$$

Then there exist disjoint open sets U and V in M such that M is the union of U , V , and their common boundary; and such that

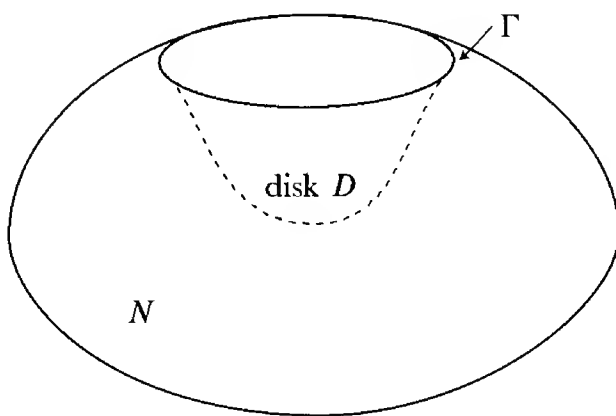
- (1) $K \geq 0$ on U and $K \leq 0$ on V .
- (2) $i: U \rightarrow i(U)$ is a diffeomorphism, and $i(U)$ is an open subset of the set $\partial H(i(M))$, bounded by a finite number of convex plane curves, each plane being the common tangent plane of $i(M)$ along the curve.

PROOF. Consider a topset of $i(M)$: it is of the form $P \cap M$ for some support plane P of $H(i(M))$, and $C = P \cap H(i(M))$ is a convex set. If this convex set contains only one point, then this point is in $i(M)$. If the convex set is a line

segment, then the endpoints must be in $i(M)$, so the whole segment must be in $i(M)$, by the first part of Lemma 15. Otherwise, the convex set is bounded by a curve Γ in P . We claim that all points of Γ are in $i(M)$. This is clear for all points of Γ which are extreme points of C (points which are not between other points of C). So the only possible exceptions are points on straight line segments of Γ . But the endpoints of such segments must be in $i(M)$, and then the whole segment must, by the second part of Lemma 15.



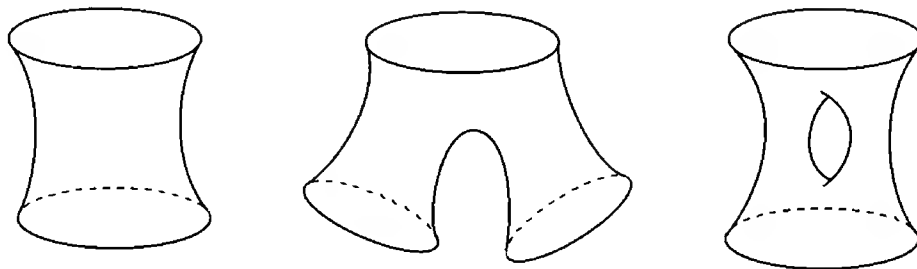
It is possible that Γ bounds a disc on $i(M)$, for the whole interior of Γ in P may be part of $i(M)$. But it is not possible that Γ is the boundary of a disc with $K \leq 0$ everywhere and $K < 0$ somewhere. For suppose it were. The disc D would be tangent to P along Γ . Let N be a surface tangent to P along Γ



such that the union of N and the disc which Γ bounds in P is convex. Then $N' = N \cup D$ is homeomorphic to S^2 , but

$$\begin{aligned} \int_{N'} K \, dA &= \int_N K \, dA + \int_D K \, dA \\ &= 4\pi + \int_D K \, dA \\ &< 4\pi, \end{aligned}$$

a contradiction. So if the interior of Γ is not part of $i(M)$, then Γ must be joined to a similar curve Γ' , or to several such curves. Since M is compact,

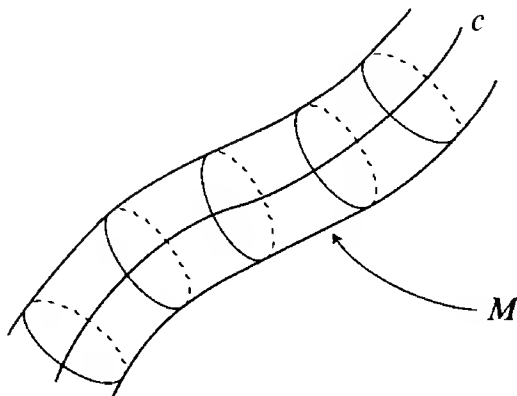


there can be only finitely many such curves $\Gamma_1, \dots, \Gamma_k$. Then $H(i(M))$ minus the discs bounded by $\Gamma_1, \dots, \Gamma_k$ can be taken to be $i(U)$. ♦

The study of total curvature (for submanifolds of Euclidean spaces in general) has recently grown into a little field of its own. However, almost all the results require methods from topology or Morse Theory, which are beyond our reach. We will end instead with a few somewhat older results concerning immersions $i: S^1 \rightarrow \mathbb{R}^3$, or equivalently, closed curves $c: [a, b] \rightarrow \mathbb{R}^3$. If c is parameterized by arclength, we define the **total curvature** of c to be

$$\int_a^b \kappa(s) ds = \int_a^b |\mathbf{t}(s)| ds = \text{length of the curve } \mathbf{t}: [a, b] \rightarrow S^2,$$

where $\kappa \geq 0$ is the ordinary curvature, and \mathbf{t} is the unit tangent vector. [This definition should not be confused with that given on pg. II.18, where we were concerned only with plane curves, and consequently allowed κ to be both positive and negative.] This total curvature can be related to the total absolute curvature of a certain surface M in \mathbb{R}^3 , the “canal surface” formed from the union of circles which bound discs of radius ε perpendicular to c (the Adden-

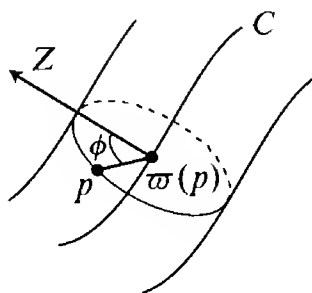


dum to Chapter I.9 can be used to show that M is an immersed surface for sufficiently small ε).

17. PROPOSITION. If M is a canal surface of the closed curve $c: [a, b] \rightarrow \mathbb{R}^3$, then

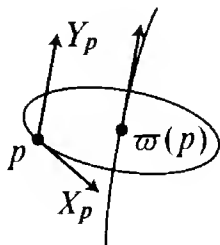
$$\int_a^b \kappa(s) ds = \frac{1}{4} \int_M |K| dA = \frac{1}{2} \int_{\{p \in M: K(p) > 0\}} K dA.$$

PROOF. Let $C = c([a, b])$ and $\varpi: M \rightarrow C$ the projection with $\varpi(p) = q$ when $p \in M$ is on the circle of radius ε perpendicular to C at q . Choose a unit vector field Z along C which is everywhere normal to C . On $M - \varepsilon \cdot Z$ we can



define a function ϕ , with values in $(0, 2\pi)$, giving the angle between p and Z on the circle $\varpi^{-1}(\varpi(p))$. If we regard the arclength s as a function on C (or $C - c(a)$, to be more rigorous), and also let s denote the function $s \circ \varpi$ on M , then (ϕ, s) is a coordinate system on M (minus a set of measure 0).

For each $p \in M$, let X_p be the unit vector tangent to the circle $\varpi^{-1}(\varpi(p))$ at p , and let Y_p be the unit vector at p which is parallel to the tangent vector dc/ds at $\varpi(p)$. The normal $N(p)$ at p points in the direction from $\varpi(p)$ to p



(Problem 2), so X_p and Y_p are tangent to M at p . It is easy to see that

$$\left. \begin{aligned} d\phi(X) &= 1 \\ ds(X) &= 0 \\ ds(Y) &= 1 \end{aligned} \right\} \implies d\phi \wedge ds(X, Y) = 1.$$

This shows that

$$d\phi \wedge ds = dA, \quad \text{the volume element on } M.$$

So

$$\int_M |K| dA = \int_M |K| d\phi \wedge ds = \int_a^b \left(\int_0^{2\pi} |K(\phi, s)| d\phi \right) ds.$$

Now

$$dN(X_p) = X_p, \quad \text{since } N \text{ is the identity map on the circle } \varpi^{-1}(\varpi(p)),$$

so if p has coordinates (ϕ, s) , then

$$\begin{aligned} \text{(I)} \quad K(\phi, s) &= -\langle dN(Y_p), Y_p \rangle = -\langle \nabla'_{Y_p} N, Y_p \rangle \\ &= \langle N(\phi, s), \nabla'_{Y_p} Y \rangle \quad \text{since } Y \text{ is a unit vector field} \\ &= \langle N(\phi, s), \kappa(s) \cdot \mathbf{n}(s) \rangle \\ &= \kappa(s) \cdot \{\text{cosine of the angle between } N(\phi, s) \text{ and } \mathbf{n}(s)\} \\ &= \kappa(s) \cdot \cos(\phi - \phi_0), \text{ say.} \end{aligned}$$

(When $\kappa(s) = 0$ and $\mathbf{n}(s)$ is undefined, this formula still holds, with an arbitrary choice of ϕ_0). Thus we have

$$\int_0^{2\pi} |K(\phi, s)| d\phi = \kappa(s) \int_0^{2\pi} |\cos(\phi - \phi_0)| d\phi = 4\kappa(s),$$

which shows that

$$\int_M |K| dA = \int_a^b 4\kappa(s) ds.$$

The second equality can easily be deduced from the fact that

$$\int_M K dA = 0,$$

which follows from the Gauss-Bonnet Theorem, since M is a torus. We can also deduce the result directly from equation (I), for if $\kappa(s) \neq 0$, then $K(\phi, s) \geq 0$ for $\cos(\phi - \phi_0) \geq 0$, and the integral of $\cos(\phi - \phi_0)$ over the subset of $[0, 2\pi]$ where it is ≥ 0 is exactly 2. ♦

18. COROLLARY (FENCHEL). The total curvature of any closed curve $c: [a, b] \rightarrow \mathbb{R}^3$ is $\geq 2\pi$. Equality holds if and only if c is a plane convex curve.

PROOF. The inequality follows directly from Theorem 13 and Proposition 17.

Now suppose that equality holds. In the proof of Theorem 13 we actually showed that

$$\int_{\tilde{M}} K dA \geq 4\pi,$$

where \tilde{M} is the set of all points $p \in M$ such that M lies on one side of M_p . We clearly also have

$$\int_{\tilde{\tilde{M}}} K dA \geq 4\pi,$$

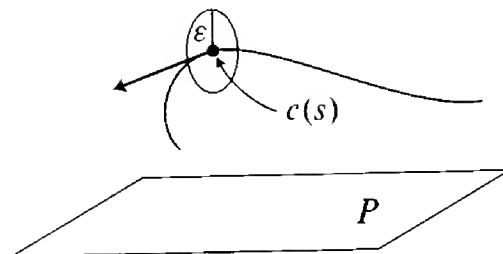
where $\tilde{\tilde{M}} = \tilde{M} \cap \{p \in M : K(p) > 0\}$. Now if $\{p \in M : K(p) > 0\} - \tilde{\tilde{M}} \neq \emptyset$, then we would have

$$\int_{\{p \in M : K(p) > 0\}} K dA > 4\pi,$$

since $\tilde{\tilde{M}}$ is a closed subset of $\{p \in M : K(p) > 0\}$. This would contradict the assumption that the total curvature of c is 2π , by Proposition 17. So we see that

$$K(p) > 0 \implies M \text{ lies on one side of } M_p.$$

Now consider a point $c(s)$ with $\kappa(s) \neq 0$. On the circle $\varpi^{-1}(c(s))$ there is an open semi-circle with $K > 0$. At the two endpoints of this semi-circle the tangent planes P_1, P_2 of M are parallel. Moreover, M lies on one side of each of these planes, since the points are limits of points where $K > 0$, and hence of points p such that M lies on one side of M_p . So M lies entirely within the space between the parallel planes P_1, P_2 . These planes are at distance 2ε if M is the canal surface formed by discs of radius ε . We claim that C must lie entirely on the plane P midway between P_1 and P_2 . For consider a point $c(s)$ of C furthest



from P . The tangent vector $c'(s)$ must be parallel to P , which means that the

disc of radius ε through $c(s)$ and perpendicular to c is also perpendicular to P . But then some points of this disc will lie outside of the region bounded by P_1 and P_2 unless $c(s)$ is on P . Thus c is a plane curve. The proof that c is convex is left to the reader. ♦

This proof, due to Voss [1], is quite different from the original proof of Fenchel [1]. For the sake of completeness, we offer the following extremely simple proof of Horn [1], where references to many other proofs may be found.

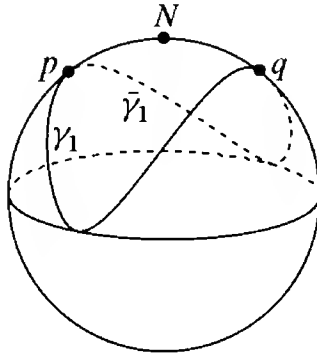
SECOND PROOF. We claim that if the closed curve $\mathbf{t}: [a, b] \rightarrow S^2$ lies in a closed hemisphere of S^2 , then c must be a plane curve (this implies, moreover, that \mathbf{t} cannot lie in an open hemisphere). Indeed, suppose that \mathbf{t} lies in a closed hemisphere; without loss of generality we can assume that it is the northern hemisphere, so that the third component \mathbf{t}^3 of \mathbf{t} satisfies $\mathbf{t}^3 \geq 0$. Then

$$0 = c^3(b) - c^3(a) = \int_a^b \mathbf{t}^3(s) ds \geq 0,$$

which implies that $\mathbf{t}^3(s) = 0$ for all s , and hence c is a plane curve. We now appeal to a simple

19. LEMMA. If γ is a closed curve on S^2 of length $< 2\pi$, then γ is contained in some open hemisphere of S^2 ; if γ has length 2π , then γ is contained in some closed hemisphere.

PROOF. Choose points p, q on γ which divide it into two arcs, γ_1 and γ_2 , of equal length. Rotate γ so that the north pole $N = (0, 0, 1)$ lies on the midpoint of the shorter arc of the great circle joining p and q . If arc γ_1 intersects the



equator at some point, let $\bar{\gamma}_1$ be the arc from p to q which is symmetric to γ_1 with respect to N . Then the closed curve made up of γ_1 and $\bar{\gamma}_1$ has the same length as γ , and also *contains two antipodal points*. This shows that the length of γ

is $\geq 2\pi$, and strict inequality holds if γ_1 actually enters the southern hemisphere. Since the same considerations hold for the other arc γ_2 , the lemma is proved, and with it the Theorem. ♦

Our last result concerns knotted closed curves $c: [a, b] \rightarrow \mathbb{R}$. There are several possible ways to define when a closed curve is knotted; for our purposes it will be simplest to say that c is **unknotted** if $c([a, b])$ is the boundary of an imbedded disc, and **knotted** otherwise. The following closed curve is knotted, although proving this fact requires considerable work.



20. THEOREM (FARY, MILNOR). The total curvature of a knotted closed curve $c: [a, b] \rightarrow \mathbb{R}^3$ is $\geq 4\pi$.

PROOF. Suppose that the total curvature of c is $< 4\pi$. If M is a canal surface of c , then by Proposition 17

$$\int_M |K| dA < 16\pi \implies \int_{S^2} \# da < 16\pi,$$

where da is the volume element of S^2 . This means that some point $v \in S^2$ is the image of at most 3 points of M , which is equivalent to saying that $c'(s)$ is perpendicular to v for at most 3 values of s . For simplicity, say that $v = (0, 0, 1)$. Since

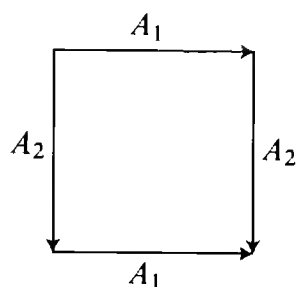
$$\langle v, c'(s) \rangle = \frac{d}{ds} \langle v, c(s) \rangle,$$

the function $s \mapsto \langle v, c(s) \rangle = (\text{height of } c(s) \text{ above } (x, y)\text{-plane})$ has derivative 0 for at most 3 values of s . Since the number of relative maxima or minima of the height function is even, there must be just one of each. Therefore the curve c must consist of two arcs joining the lowest and highest points, each arc having monotonically increasing height. Each plane parallel to the (x, y) -plane between these lowest and highest points intersects the curve in 2 points. Joining each such pair by the line segment between them, we obtain an imbedded disc whose boundary is the curve. ♦

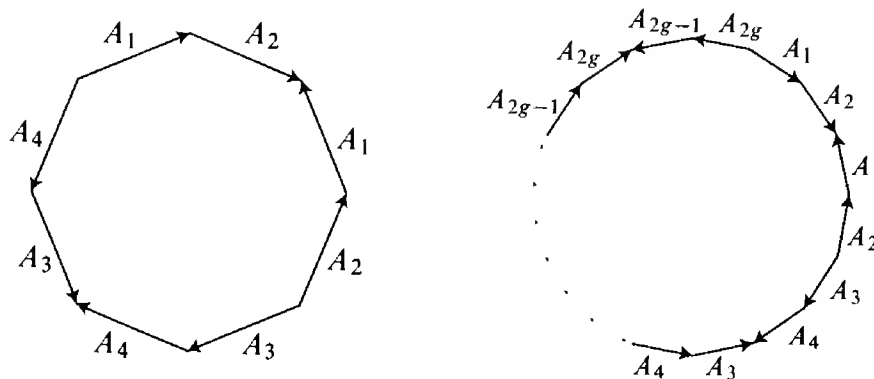
ADDENDUM 1

COMPACT SURFACES WITH
CONSTANT NEGATIVE CURVATURE

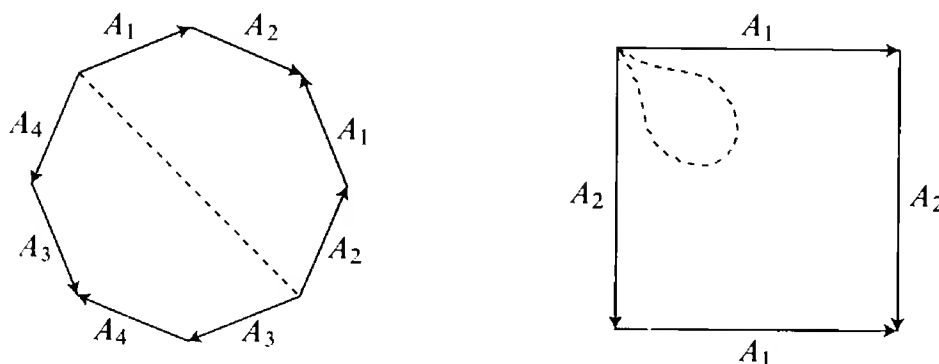
In order to construct a Riemannian metric $\langle \cdot, \cdot \rangle$ with $K = -1$ on any compact surface M of genus $g \geq 2$, we first need to consider the standard topological way of obtaining these surfaces. For simplicity we will work only with the oriented ones. We have often used the fact that the torus, with genus $g = 1$, can be obtained by identifying sides of a square according to the scheme shown below.



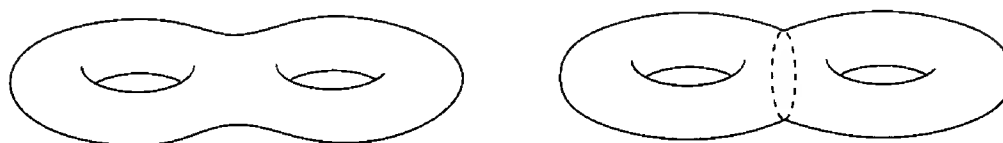
In general, the “ g -holed torus”, with genus $g \geq 1$, can be obtained by identifying sides of a $4g$ -gon according to the following scheme:



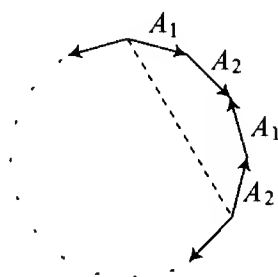
If your powers of visualization are much better than mine, you may be able to literally see that this is the case. Otherwise, the following argument should convince you. The dashed line in the first figure below is a circle, because, as



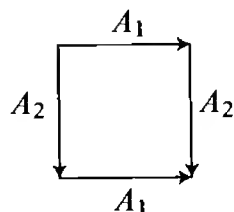
a quick check shows, the required identifications of sides forces all vertices to be identified to one point. This circle divides the identification space into two parts. The right half of the figure shows that each part, together with the bounding circle, is homeomorphic to a torus with an open disc removed. So the identification space is homeomorphic to the space obtained by removing a disc from each of two tori, and then identifying the boundary circles; hence



it is a 2-holed torus. A similar argument can be used to treat the general case, by induction.



Now the flat metric on the torus is just obtained from the flat metric on the square by performing the required identifications. The fact that we actually get



a Riemannian metric on the identification space depends on two circumstances: first, the opposite sides are of equal length, and second, the sum of the angles at the four vertices is exactly 2π . It is the failure of this second condition in larger polygons which prevents us from getting a flat metric on the orientable surfaces of higher genus. For surfaces with $g > 1$ we will construct a metric with $K = -1$ by obtaining $4g$ -gons in the *non-Euclidean plane* whose angles add up to exactly 4π .

Our model for the non-Euclidean plane will be the Poincaré upper half-plane $\mathcal{H}^2 = \{(x, y) \in \mathbb{R}^2 : y > 0\}$ with the Riemannian metric

$$\langle \cdot, \cdot \rangle = \frac{dx \otimes dx + dy \otimes dy}{y^2};$$

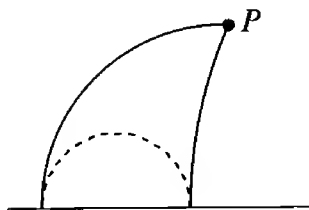
this manifold has constant curvature $K = -1$ (compare Problem I.9-41 and

pg. II.301). It would be possible to check that if we define the “straight lines” of \mathcal{H}^2 to be the geodesics for $(\ , \)$, then the Poincaré upper half-plane satisfies the axioms for non-Euclidean geometry, and then conclude, by a theorem of non-Euclidean geometry, that the sum of the angles of a geodesic triangle is always $< \pi$. Happily, we can also reach this conclusion simply by applying the Gauss-Bonnet Formula. Indeed we find (Corollary 7) that for a geodesic triangle Δ with interior angles $\iota_1, \iota_2, \iota_3$ we have

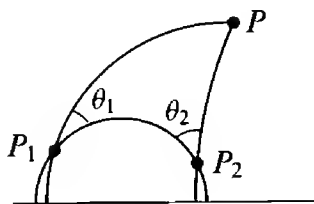
$$\pi - (\iota_1 + \iota_2 + \iota_3) = - \int_{\Delta} -1 dA = \text{area}(\Delta) > 0.$$

This also shows us that small triangles are very close to Euclidean triangles: the sum $\iota_1 + \iota_2 + \iota_3$ can be made as close to π as we like. It follows that the interior angles of an n -gon can be made as close to $(n - 1)\pi$ as we like.

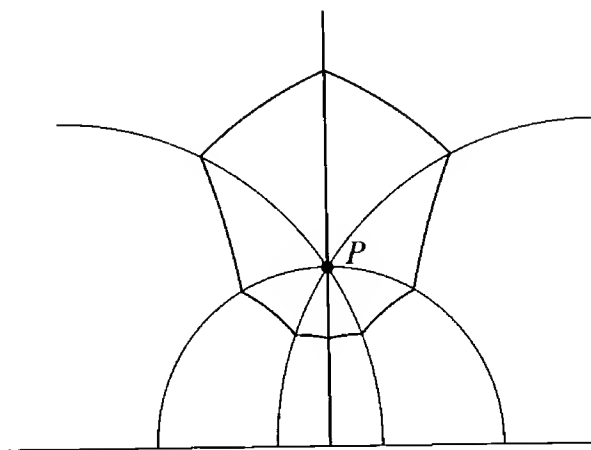
In contrast to the situation for small triangles, let us consider what happens when we take 2 fixed geodesic rays from a point P and join points far out on each of the two sides with a third geodesic. The figure below shows two geodesic rays, both of which are portions of circles or straight lines which meet the x -axis at right angles (Problem I.9-41). The dashed line is another geodesic



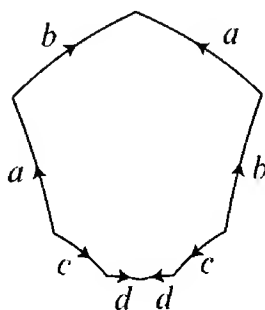
which “meets these at 0 angle”. Of course, this third geodesic doesn’t really meet either of the first two in \mathcal{H}^2 . But it is clear from this picture that if we take points P_1 and P_2 far enough out on the two original geodesics, then the angles θ_1 and θ_2 between the unique geodesic through P_1 and P_2 and our given geodesics can be made as small as we like.



Now let us take a point P in \mathcal{H}^2 , and draw $4g$ geodesic rays from P , the angle between two successive rays being $2\pi/4g$. For each $r > 0$, consider the equilateral $4g$ -sided geodesic polygon obtained by joining the points on these rays which are at distance r from P . Let $\Sigma(r)$ be the sum of the interior angles



of this polygon; clearly $\Sigma(r)$ is a continuous function of r . But we have seen that $\Sigma(r) \rightarrow 0$ as $r \rightarrow \infty$, while as $r \rightarrow 0$ we have $\Sigma(r) \rightarrow (4g - 1)\pi > 2\pi$ for $g \geq 2$. So if $g \geq 2$, then there is an r with $\Sigma(r) = 2\pi$. This gives us an equilateral $4g$ -gon with the sum of the interior angles $= 2\pi$. After identifying



pairs of sides according to the scheme on page 292, we then have a metric on the surface of genus g with constant curvature $K = -1$.

We add here some supplementary remarks which may be of interest to readers familiar with complex analysis. The necessary identification of pairs of sides can easily be accomplished by means of one-one complex analytic maps of \mathcal{H}^2 onto itself, and it is thus easy to see that M can be made into a Riemann surface (a complex manifold of complex dimension 1), meaning that there is a collec-

tion \mathcal{A} of homeomorphisms $f: U \rightarrow \mathbb{C} = \mathbb{R}^2$, from open subsets $U \subset M$ onto open subsets of \mathbb{C} , such that

- (i) the union of the domains of all $f \in \mathcal{A}$ covers M
- (ii) if $f: U \rightarrow \mathbb{C}$ and $g: V \rightarrow \mathbb{C}$ are in \mathcal{A} , then

$$g \circ f^{-1}: f(U \cap V) \rightarrow g(U \cap V)$$

is complex analytic.

[In fact, it is possible to establish even more. Since each vertex of our polygon has angle $2\pi/4g$, we can arrange exactly $4g$ polygons congruent to it around every vertex. The same construction can be carried out at the new vertices of the new polygons, and the construction can then be repeated indefinitely. In this way we arrive at a “tiling” of the hyperbolic plane by equilateral $4g$ -gons. The one-one complex analytic maps of \mathcal{H}^2 onto itself which are required for identifying the sides of our original polygon all preserve the tiling, and generate a group \mathcal{G} , with M homeomorphic to the quotient space $\mathcal{H}^2/\mathcal{G}$ obtained by identifying z and $g(z)$ for all $z \in \mathcal{H}^2$ and $g \in \mathcal{G}$. The map $\mathcal{H}^2 \rightarrow \mathcal{H}^2/\mathcal{G}$ defined by taking z into its equivalence class in $\mathcal{H}^2/\mathcal{G}$ is a covering map. Thus we obtain an explicit construction of a covering map $\pi: \mathcal{H}^2 \rightarrow M$, which shows that \mathcal{G} must be isomorphic to $\pi_1(M)$, and that the universal covering space of M is homeomorphic to \mathbb{R}^2 . (The latter fact could also have been deduced from purely topological considerations, since all simply-connected (paracompact) 2-manifolds are homeomorphic to either \mathbb{R}^2 or S^2 , and S^2 cannot be the universal covering space of M since S^2 is compact, while $\pi_1(M)$ is infinite.)]

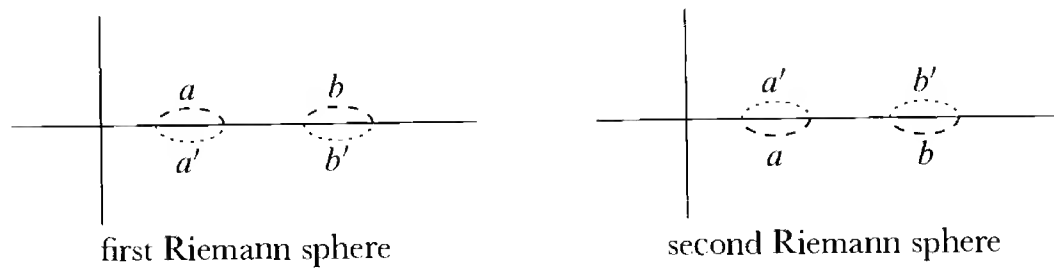
There are also two other methods which we can use to put a complex manifold structure on M .

Method A. Choose an arbitrary Riemannian metric $\langle \cdot, \cdot \rangle$ for M . We recall (pg. II.296) that a map f between Riemannian manifolds is **conformal** if each f_* is angle preserving. For any point p of a 2-dimensional Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, there is a neighborhood U of p and a conformal diffeomorphism $f: U \rightarrow \mathbb{R}^2$ onto an open subset of \mathbb{R}^2 with its usual Riemannian metric; this fact was mentioned in Volume II, and a proof will be found in Addendum 1 to Chapter 9 of these Volumes. It is also an elementary fact (Problem 4-9) that a diffeomorphism $f: W \rightarrow \mathbb{C}$, from an open set $W \subset \mathbb{C}$ onto an open subset of \mathbb{C} , is complex analytic if and only if f is orientation preserving and conformal with respect to the usual metric on \mathbb{R}^2 . We can therefore define \mathcal{A} to be the collection of all conformal orientation preserving $f: U \rightarrow \mathbb{R}^2$; if $f, g \in \mathcal{A}$, then $g \circ f^{-1}$ is orientation preserving and conformal, so it is complex analytic.

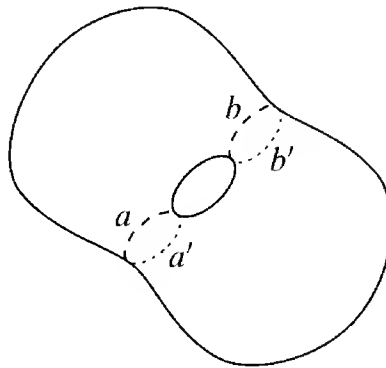
Method B. This method involves the Riemann surface, including branch points, of a “complete analytic function”. From its definition, it is clear that this surface is a Riemann surface in the sense of being a complex manifold. On the other hand, the usual method for visualizing the Riemann surface of

$$\sqrt{(z-1)(z-2)\dots(z-2(g+1))}$$

is to take two copies of the Riemann sphere, make “cuts” from 1 to 2, from 3 to 4, \dots , and from $2g+1$ to $2g+2$, and identify the corresponding cuts in the



two different spheres. This is homeomorphic to the g -holed torus.



If we use one of these two methods for putting a complex manifold structure on a compact oriented surface M of genus $g \geq 2$, then we can give another construction of a metric on M with constant curvature $K = -1$, provided that we are willing to use yet more machinery. We consider the universal covering space $\pi: \tilde{M} \rightarrow M$, and give it the structure of a Riemann surface in the obvious way. Then M is the quotient of \tilde{M} by the group of covering transformations, each of which is a one-one complex analytic map of \tilde{M} onto itself. Now we expect that \tilde{M} is \mathcal{H}^2 (as we have already mentioned at the beginning of this whole discussion). We can establish this fact independently by using the “general uniformization theorem”, which tells us that the simply-connected Rie-

mann surface \tilde{M} must be analytically equivalent to either \mathbb{C} or the upper half-plane \mathcal{H}^2 . Now all one-one complex analytic maps of \mathbb{C} onto itself are of the form $z \mapsto az+b$. This group, and hence any subgroup, is abelian, while the fundamental group of M is non-abelian if it has genus > 2 . So if M is a compact surface of genus $g \geq 2$, then \tilde{M} is \mathcal{H}^2 . But the only one-one complex analytic maps of \mathcal{H}^2 onto itself are of the form $z \mapsto (az+b)/(cz+d)$, and (Problem I.9-47, or Problem 7-6) these maps are isometries of \mathcal{H}^2 with the metric $(dx \otimes dx + dy \otimes dy)/y^2$. Consequently, M has a metric $\langle \ , \ \rangle$ such that $\pi^*\langle \ , \ \rangle = (dx \otimes dx + dy \otimes dy)/y^2$; clearly $(M, \langle \ , \ \rangle)$ has constant curvature -1 .

ADDENDUM 2

THE DEGREE OF THE NORMAL MAP

Let $M \subset \mathbb{R}^m$ be a compact hypersurface with normal map $\nu: M \rightarrow S^{m-1}$; we want to show that the degree of ν is $\chi(M)/2$. We will actually prove a more general result, involving any compact orientable manifold $M^n \subset \mathbb{R}^m$. From the Addendum to Chapter I.9 we know that for sufficiently small $\varepsilon > 0$, the set $N = \{q \in \mathbb{R}^m : d(q, M) \leq \varepsilon\}$ is a compact m -dimensional manifold-with-boundary, which is a tubular neighborhood under a projection map $\pi: N \rightarrow M$. We give N the usual orientation and ∂N the induced orientation, so that the corresponding normal map $\nu: \partial N \rightarrow S^{m-1}$ is outward pointing. We will show that:

The degree of the normal map $\nu: \partial N \rightarrow S^{m-1}$ is $\chi(M)$.

[In the special case where $M \subset \mathbb{R}^m$ is a hypersurface, the manifold ∂N consists of two components each homeomorphic to M , and the degree of $\nu: \partial N \rightarrow S^{m-1}$ is just twice the degree of the normal map of M , so it will follow that this degree is $\chi(M)/2$.]

We will follow the exposition in Milnor [1]. The first step is a simple

21. LEMMA. Let $N \subset \mathbb{R}^m$ be a compact m -dimensional manifold-with-boundary, and let $\nu: \partial N \rightarrow S^{m-1}$ be the normal map. Let X be a vector field on N with isolated zeros, and suppose that X is outward pointing on ∂N . Then the degree of $\nu: \partial N \rightarrow S^{m-1}$ is equal to the sum of the indices of X .

PROOF. We regard X as a map $X: N \rightarrow \mathbb{R}^m$. Let p_1, \dots, p_k be the zeros of X , and let U_1, \dots, U_k be open ε -balls around p_1, \dots, p_k with all $\overline{U_i} \subset \text{interior } N$. Then $\mathcal{B} = N - (U_1 \cup \dots \cup U_k)$ is a compact manifold-with-boundary, and $\bar{X} = X/|X|: \mathcal{B} \rightarrow S^{m-1}$. Now if ω is any $(m-1)$ -form on S^{m-1} , then

$$\int_{\partial \mathcal{B}} \bar{X}^*(\omega) = (\text{degree } \bar{X}|_{\partial \mathcal{B}}) \cdot \int_{S^{m-1}} \omega.$$

By Stokes' Theorem we have

$$\int_{\partial \mathcal{B}} \bar{X}^*(\omega) = \int_{\mathcal{B}} d\bar{X}^*(\omega) = \int_{\mathcal{B}} \bar{X}^*(d\omega) = 0,$$

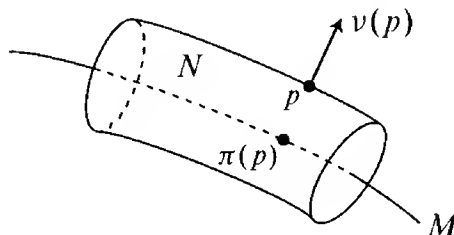
so the degree of $\bar{X}|_{\partial \mathcal{B}}$ must be 0. But $\bar{X}|_{\partial N}$ is smoothly homotopic to ν , while the degrees on the other components of $\partial \mathcal{B}$ are the negatives of the indices of X at the p_i (the minus signs come from the fact that the orientations on these components are the negatives of the usual ones). Thus

$$\deg \nu - (\text{sum of indices of } X \text{ at the } p_i) = 0. \quad \blacklozenge$$

Now suppose we have a compact oriented manifold $M^n \subset \mathbb{R}^m$ and we choose an arbitrary vector field X on M with only isolated zeros; according to the Poincaré-Hopf Theorem (I.11-30) the sum of its indices is $\chi(M)$. We can extend X to a vector field Y on the tubular neighborhood N in a rather obvious way, by defining

$$(*) \quad Y(p) = [p - \pi(p)] + X(\pi(p)) \quad (\text{considered as a vector at } p).$$

Since the normal $\nu(p)$ points in the direction from $\pi(p)$ to p (Problem 2), and the vector $p - \pi(p)$ is perpendicular to M_p , it is clear that Y points outward



along ∂N , and that the zeros of Y in all of N are precisely the zeros of X in M . If we could show that the index of Y at such a zero equals the index of X , then the desired result would follow from Lemma 21. However, a direct analysis of the index of Y turns out to be very difficult, so we take a slight detour.

Consider first a vector field X on \mathbb{R}^n , with an isolated zero at $0 \in \mathbb{R}^n$. We regard X as a function $X: \mathbb{R}^n \rightarrow \mathbb{R}^n$, and we define X to be **non-degenerate** at 0 if the derivative $DX(0): \mathbb{R}^n \rightarrow \mathbb{R}^n$ is non-singular.

22. LEMMA. If X has a non-degenerate zero at 0, then the index of X at 0 is $+1$ or -1 , depending on whether $\det DX(p) > 0$ or < 0 .

PROOF. We can assume $p = 0$. Then X is a diffeomorphism on some convex open neighborhood U of 0. In the proof of Lemma I.11-27 we saw that if X is orientation preserving, then X is smoothly homotopic to the identity via maps which have no zeros. So the index is 1. Similarly, if X reverses orientation, then it is smoothly homotopic to a reflection, and has index -1 . ♦

Now consider an oriented n -manifold M , and a vector field X on M with an isolated zero at $p \in M$. Let $f: U \rightarrow \mathbb{R}^n$ be an orientation preserving diffeomorphism, where $U \subset M$ is an open set containing p , and $f(p) = 0$. Then f_*X is a vector field on \mathbb{R}^n , and we define X to be **non-degenerate** at p if f_*X is non-degenerate at 0. If $M \subset \mathbb{R}^m$ is a submanifold of \mathbb{R}^m , and we regard

the vector field X on M as a function $X: M \rightarrow \mathbb{R}^m$, then $DX(p): \mathbb{R}^m \rightarrow \mathbb{R}^m$ will take M_p to M_p . In fact, for the function $X \circ f^{-1}: \mathbb{R}^n \rightarrow M$ we have $(X \circ f^{-1})_*: \mathbb{R}^n_0 \rightarrow M_p$; but $(f^{-1})_*(\mathbb{R}^n_0) = M_p$, while X_* is the same as DX on M_p . It is easy to see from Lemma 22 that if p is a non-degenerate 0, then the index of X at 0 is $+1$ or -1 , depending on whether $\det DX(p): M_p \rightarrow M_p$ is > 0 or < 0 .

23. THEOREM. Let $M^n \subset \mathbb{R}^m$ be a compact oriented manifold with a closed tubular neighborhood $\pi: N \rightarrow M$. Then the degree of the normal map $\nu: \partial N \rightarrow S^{m-1}$ is $\chi(M)$.

PROOF. We will show that there is a vector field X on M with only non-degenerate zeros. Assuming this fact for the moment, we use the vector field X to define a vector field Y on N by $(*)$. It is easy to see that for $p \in M$ we have

$$DY(p) = \begin{cases} DX(p) & \text{on } M_p \\ I & \text{on } M_p^\perp. \end{cases}$$

Therefore $\det DY(p) = \det DX(p): M_p \rightarrow M_p$. So the index of Y at a zero p equals the index of X at p . Then Lemma 21 implies that the degree of $\nu: \partial N \rightarrow S^{m-1}$ is the sum of the indices of X , which equals $\chi(M)$ by the Poincaré-Hopf Theorem.

To obtain the desired vector field on M , we first choose a vector field X with only finitely many (possibly degenerate) zeros (as on pp. I.449–450). It obviously suffices to show that for each zero p we can find a new vector field which equals X outside of a small neighborhood U of p , and has only non-degenerate zeros inside the neighborhood. It obviously suffices to work on \mathbb{R}^n . Given a neighborhood U of a zero $p \in \mathbb{R}^n$, let $f: \mathbb{R}^n \rightarrow [0, 1]$ be a C^∞ function which is 1 on an open set W with $p \in W \subset \overline{W} \subset U$, and 0 on $\mathbb{R}^n - U$. By Sard's Theorem, there is a regular value X_0 of X arbitrarily close to 0. Consider the vector field

$$\bar{X} = X - f \cdot X_0.$$

Within W this vector field is 0 only at points q with $X(q) = X_0$, so all zeros in W are non-degenerate; on the other hand, in $U - W$ there are no zeros at all if X_0 is sufficiently small. ♦

As a final remark, we note that if we choose U to be a closed manifold-with-boundary, then the sum of the indices of \bar{X} at zeros within U is just the degree of $\bar{X}/|\bar{X}|: \partial U \rightarrow S^{m-1}$, as in the proof of Lemma 21. But this map is

the map $X/|X|: \partial U \rightarrow S^{m-1}$, whose degree is, by definition, the index of X at p . So, without using the Poincaré-Hopf Theorem, we have reproved the fact that for any vector field X on M with only isolated zeros, the sum of its indices is a constant, namely $\deg v: \partial N \rightarrow S^{m-1}$. We could then identify this constant with $\chi(M)$ as we did in Chapter I.11, when we originally proved the Poincaré-Hopf Theorem.

PROBLEMS

1. Let $X \subset \mathbb{R}^m$ be any set.
 - (a) The convex hull $H(X)$ is the union of the convex hulls of all finite subsets of X .
 - (b) The convex hull of $m + 2$ points in \mathbb{R}^m is the union of the convex hulls of all its subsets of $m + 1$ points.
 - (c) If X is compact, then so is $H(X)$.
2. Prove the assertions about the normals to a canal surface on page 287, and the normals of the tubular neighborhood, on page 300, by using the argument in Problem 3-12. Also generalize the argument of Problem I.9-28.

MINI-BIBLIOGRAPHY FOR VOLUME III

Brackets [] indicate journal articles, braces { } indicate books.

Bol, G.

- [1] *Über Nabelpunkte auf einer Eifläche*, Math. Z. **49** (1944), 389–410.

Efimov, N. V.

- [1] *Generation of singularities on surfaces of negative curvature* (Russian), Mat. Sb. **64** (1964), 286–320.

Fenchel, W.

- [1] *Über Krümmung und Windung geschlossener Raumkurven*, Math. Ann. **101** (1929), 238–252.

Hamburger, H.

- [1] *Beweis einer Carathéodoryschen Vermutung. Teil I*, Ann. of Math. **41** (1940), 63–86; *II*, Acta Math. **73** (1941), 175–228; *III*, Acta Math. **73** (1941), 229–332.

Hilbert, D.

- [1] *Ueber Flächen von constanter Gausscher Krümmung*, Trans. Amer. Math. Soc. **2** (1901), 87–99.

Holmgren, E.

- [1] *Sur les surfaces à courbure constante négative*, C. R. Acad. Sci. Paris Ser. A-B **134** (1902), 740–743.

Horn, R. A.

- [1] *On Fenchel's theorem*, Amer. Math. Monthly **78** (1971), 380–381.

Klotz, T.

- [1] *On G. Bol's Proof of Carathéodory's conjecture*, Comm. Pure Appl. Math. **12** (1959), 277–311.

Klotz Milnor, T.

- [2] *Efimov's Theorem about complete immersed surfaces of negative curvature*, Advances in Math. **8** (1972), 474–543.

McCleary, J.

- {1} *On Jacobi's remarkable curve theorem*, Historia Math. **21** (1994), 377–385.

Milnor, J. W.

- {1} *Topology from the Differentiable Viewpoint*, University Press of Virginia, Charlottesville, Virginia, 1965.

Voss, K.

- [1] *Eine Bemerkung über die Totalkrümmung geschlossener Raumkurven*, Arch. Math. (Basel) **6** (1955), 259–263.

Wunderlich, W.

- [1] *Über ein abwickelbares Möbiusband*, Monatsch. Math. **66** (1962), 276–289.

NOTATION INDEX

CHAPTER 1

$K(P)$	5	\mathcal{K}_{ext}	128
$K'(P)$	5	k_1, k_2	48
M_p^\perp	1	$\mathbf{k}_1, \mathbf{k}_2$	128
s	4, 19	$\mathcal{L}_{\mu k}$	127
s_{ij}^r	19	l	34
θ^i	16	l_{ij}	34
ν	7	l_i^h	53
$\nu^\beta{}_{;\rho}$	14	ℓ_{ijk}	105
ϕ^α	17	ℓ_{ij}^k	105
Ψ_β^α	17	m	34
ψ_β^α	17	N	33
Ω_j^i	16	N_f	33
ω_j^i	16	\mathcal{N}	101
∇'	2	n	34
Π	8	\mathcal{R}	123
\mathbb{T}	1	S	77, 80
\perp	1	\mathcal{S}	126
		$\tilde{\mathcal{S}}$	127
		$\mathcal{S} * \mathbb{I}$	130
		s	104

CHAPTER 2

b_i^j	122	\mathbf{t}	35
C_{ij}	127	α_f	72
c_{ijk}	106	Γ_{ij}^k	108
dA	69	v_p	101
d_{ij}	84	v_*	102
$d\mathbf{v}$	102	$\Phi^{\mathbf{X}}$	37, 91
E	31	(Φ, Θ)	94
F	31	(Φ, Ψ)	95
G	31	$\Psi^{\mathbf{X}}$	91
g_{ij}	32	ω_i^j	106
g_{ij}	83, 89	ω	71
H	49	ω^i	71
\mathcal{H}	128	\mathbf{I}_f	32
$h^{\mathbf{X}}$	36, 75	\mathbf{II}_f	34
\tilde{h}	76	\mathbf{III}	62
J	116	\mathbf{III}_f	62
K	49	\mathbf{IV}	62
\mathcal{K}	125	\mathbf{I}	82, 89
		\mathbf{I}_f	83, 89

\mathbb{I}	105	CHAPTER 4	
\mathbb{I}_f	105	\bar{X}	189
$\langle \ , \ \rangle$	94, 116	$\mathbf{u}(s)$	187
$\langle \ , \ \rangle$	82, 89	$\mathbf{v}(s)$	191
$\langle \ , \ \rangle_p$	82, 89	$\kappa_g(s)$	187
$\langle \ , \ \rangle_p$	116	$\kappa_n(s)$	187
\mathbb{T}	103	$\kappa_n(X)$	188
\perp	103	τ_g	191
∇	104	$\tau_g(X)$	191
$;$	109	ϕ	189
$[ij, k]$	110	$\phi(s)$	189
CHAPTER 3		CHAPTER 6	
h	184	$H(X)$	281
p	184	δ_i	268
		ι_i	267
		$\#(p)$	278

INDEX

- Adapted moving frame, 17
- Affine, *see also* Special affine
 - conformal structure, 79
 - invariants, 71
 - map, 71
 - normal direction, 97
 - special linear, 71
 - special orthogonal, 71
- Ambient space, 1
- Analytic flat Möbius strip, 239
- Angle
 - between two vectors, 262
 - interior, 267
- Apolar, 94, 95
- Apolarity condition(s), 95, 107
 - geometric interpretation of, 115
- Approximate a surface up to second order, 36
- Asymptote, 49
- Asymptotic
 - at a point, 196
 - curve, 195
 - directions, 49
 - Tschebyscheff net, 251
 - vector, 136
- Auto-parallel, 22
- Axes, principal, 48

- Beltrami, E., 200, 226
- Beltrami-Enneper Theorem, 200
- Bol, G., 199
- Bonnet, O., 56, 185, 203
- Branch points, 221, 297
- Branched covering space, 221

- Canal surface, 286
- Canonical parameterization
 - for catenoid, 161
 - for surface of revolution, 158
- Carathéodory, C., 199
- Cartan, Élie, 193
- Cartan's Lemma, 18

- Catenary, 160
- Catenoid, 160, 170
 - canonical parameterization for, 161
- Cayley-Hamilton Theorem, 62
- Characteristic
 - line, 180
 - point, 180
- Clairaut, A. C., 214
- Classical
 - classification of developable surfaces, 237
 - counterexample, 166
 - flat surfaces, 141
 - tensor analysis treatment of submanifolds, 12
- Codazzi-Mainardi equations, 10, 11, 16, 20, 56, 70, 74, 134, 217
 - special affine, 132
- Codimension, 1
- Compact surfaces of constant negative curvature, 292
- Complete analytic function, 297
- Complete surfaces of constant curvature, 233 ff.
- Cone, generalized, 142
- Conformal, 208, 296
 - structure, 79
- Connection
 - forms, 16
 - induced, 23
- Constant curvature, 11
 - compact surfaces of negative, 292
 - complete surfaces of, 233 ff.
 - isometry of simply-connected manifolds with same, 30 ff.
 - manifolds, 25 ff.
 - rotation surfaces of, 161
- Continuation of an isometry, 30
- Convex, 64
 - hull, 281
- Covering space, branched, 221
- Cubic
 - forms, invariants of, 111
 - osculating, 48, 111
- Curvature, *see also* Constant curvature
 - absolute, total, 278
 - forms, 16

Curvature (*continued*)

- Gaussian, 49, 136
 - in orthogonal coordinate system, 217
- geodesic, 187
- line of (q, v) , 195
- mean, 49, 136
- normal, 187
- positive, 63
- principal, 48, 136
- sectional, 5
- special affine (extrinsic), 128
- special affine mean, 128
- special affine principal, 128
- total, 286
- total absolute, 278, 286
- vector
 - geodesic, 3, 187
 - normal, 187

Cuspidal edge, 143

Cuts, 297

Cylinder

- generalized, 141
- parabolic, 41

 C^∞ , *see* Smooth

Darboux, G., 190, 201, 208

Darboux frame, 191

Degenerate, *see* Non-degenerate

Degree of the normal map, 299

Developable

- surfaces, 197, 236
 - classical classification of, 237
- tangent, 142

Development, 146

Directions

- asymptotic, 49
- principal, 48, 136
 - special affine, 128

Directrix, 147

Discontinuity of angle, 268

Distribution parameter, 148

Doubly ruled surface, 153, 155

Dual 1-forms, 16

Dupin, C., 206

Dupin indicatrix, 47

Edge

- cuspidal, 143
- of regression, 143

Efimov, N. V., 260

Ellipsoid, 151

- lines of curvature on, 206
 - in a neighborhood of an umbilic, 198

umbilics on, 152

Elliptic

- hyperboloid (of one sheet), 152
- hyperboloid (of two sheets), 154
 - umbilics on, 154
- paraboloid, 39, 154
 - umbilics on, 155

point, 39, 78

Enneper, A., 200; *see also* Beltrami-Enneper theorem

Enneper's minimal surface, 174

Envelope, 176

- of one-parameter family of planes, 179

Equations of structure of $SO(3)$, 71, 73; *see also* Structural equations

Euclidean motion, proper, 51

Euler characteristic, 271

Euler's Theorem, 188

Fary, I., 291

Fenchel, W., 289, 290

First fundamental form, 31

- of a map, 32
- special affine, 82
 - of a map, 89

First structural equation, 16

Flat, 49

Möbius strip, 149, 239

surfaces, classical, 141

torus, 61

- Frame
 - adapted moving, 17
 - Darboux, 191
 - Frenet, 191
- Frenet frame, 191
- Frobenius, G., 230
- Fundamental forms, *see* First, second, third fundamental forms
- Fundamental theorem of special affine surface theory, 132
- Fundamental theorem of surface theory, 56, 73, 74
- Gauss, C. F., 53
- Gauss formulas, 4, 14, 19, 53
 - special affine, 105
- Gauss' equation, 5, 11, 16, 20, 55, 74
- Gauss' Theorema Egregium, 5, 55, 69
- Gauss-Bonnet formula, 268
- Gauss-Bonnet theorem, 271
- Gaussian curvature, 49, 136
 - in orthogonal coordinate system, 217
- General affine invariants, 71
- General uniformization theorem, 297
- Generalized
 - cone, 142
 - cylinder, 141
- Geodesic, 3, 196
 - at a point, 24
 - curvature, 187
 - curvature vector, 3, 187
 - on surface of revolution, 214
 - on torus, 230
 - torsion, 191
 - totally, 24
- Geometric interpretation of apolarity, 115
- Hadamard, J., 66
- Hamburger, H., 199
- Hazzidakis' formula, 255
- Helicoid, 150, 170
- Hilbert, D., 233, 247
- Hilbert's theorem, 247
- Holmgren, E., 247
- Hopf Umlaufsatz, 266
 - generalization of, 269
- Horn, R. A., 290
- Hull, convex, 281
- Hyperbolic
 - paraboloid, 40, 155
 - point, 40, 78
- Hyperboloid
 - elliptic (of one sheet), 152
 - elliptic (of two sheets), 154
 - umbilics on, 154
 - of revolution, 152
 - striction curve of, 183
- Hypersurface, 7
- Idiot, any, 103
- Immersion, isometric, 1
- Index of singularity of 1-dimensional distribution, 218
- Indicatrix of Dupin, 47
- Induced connection, 23
- Interior angle, 267
- Invariant
 - affine, 71
 - meaning, 60
 - under orientation preserving change of parameter, 85
 - under proper Euclidean motions, 51
- Invariants for cubic forms, 111
- Inversion, 208
- Inward, 78, 88
- Isometric immersion, 1
- Isometry
 - continuation of, 30
 - of simply-connected manifold with same curvature, 30 ff.

- Jacobi, C. G. J., 276
 Joachimsthal, F., 203
- K , in orthogonal coordinate system, 217
 Klotz, T., 199, 260
 Knotted, 291
- Laguerre, E., 193
 Levi-Civita, T., 181
 Linear affine maps, special, 71
 Lines of curvature, 195
 on an ellipsoid, 206
 in a neighborhood of an umbilic, 198
 Liouville, J., 209
- Mainardi, G., 10; *see also* Codazzi-
 Mainardi equations
 Malz, R., 243
 Mean curvature, 49, 136
 special affine, 128
 Meridian, 156
 Meusnier's Theorem, 189
 Milnor, J. W., 291, 299
 Minimal surface, 167
 Enneper's, 174
 Scherk's, 171
 Möbius, A. F., 209
 Möbius strip, 148
 analytic flat, 239
 smooth flat, 149
 Monkey saddle, 41
 Morse theory, 286
 Motion, proper Euclidean, 51
 Moving frame, adapted, 17
- Navel point, 50, 136
 Negative constant curvature, compact
 surfaces of, 292
 Neighborhood, tubular, 299
 Non-degenerate, 93, 300
 Non-Euclidean plane, 293
 Normal
 curvature, 187
 curvature vector, 187
 direction, affine, 97
 field, unit, 7
 map, degree of, 299
 projection, 1
 special affine, 101
- One-dimensional distribution, index of
 singularity of, 218
 Orthogonal affine maps, special, 71
 Orthogonal systems of surfaces, triply,
 204
 Osculating
 circle, 224
 cubic, 48, 111
 paraboloid, 42
- Parabolic
 cylinder, 41
 point, 40, 78
 Paraboloid
 elliptic, 39, 154
 umbilics on, 155
 hyperbolic, 40, 155
 osculating, 42
 Parallel along a curve, 3, 181
 Parallel surface, 185
 Parallel, on surface of revolution, 156
 Perpendicular projection, 1
 Pick invariant, 116, 231
 Planar point, 41, 78
 Plane, 141
 support, 64
 Point inward, 78, 88

- Polygon, 266
- Positive curvature, 63
- Principal
 - at a point, 196
 - axes, 48
 - curvatures, 48, 136
 - special affine, 128
 - curve, 195
 - directions, 48, 136
 - special affine, 128
 - vector, 48, 136
- Profile curve, 156
- Projection
 - normal, 1
 - perpendicular, 1
 - tangential, 1
- Proper Euclidean motion, 51
- Pseudosphere, 163

- Quadratic, 92
 - (or quadric) surface, 118, 151
- Quasi-orthonormal, 79, 88

- Radon, J., 132
- Rectifying plane, 186
- Regression, edge of, 143
- Revolution, *see* Surface of revolution
- Riemann, G. F. B., 166
- Riemann surface, 295
- Right helicoid, 150
- Rodrigues' formula, 195
- Rotation surfaces, *see* Surface of revolution
- Ruled surfaces, 146
 - doubly, 153, 155
- Rulings, 147

- Saddle, monkey, 41
- Scherk's minimal surface, 171
- Scroll, 147
- Second fundamental form, 8, 9, 33
 - of a map, 34
 - special affine, 105
 - of a map, 105
- Second order approximation of surface, 36
- Second structural equation, 16
- Sectional curvature, 5
- Similarity, 208
- Simply-connected manifolds with same
 - constant curvature, 30
- Singularity of 1-dimensional distribution, index of, 218
- Smooth flat Möbius strip, 149
- SO(3), equations of structure of, 71
- Special affine, 71
 - Codazzi-Mainardi equations, 132
 - (extrinsic) curvature, 128
 - first fundamental form, 82, 89
 - of a map, 83, 89
 - Gauss formulas, 105
 - geometry of surfaces, 71
 - mean curvature, 128
 - normal, 101
 - principal curvatures, 128
 - principal directions, 128
 - second fundamental form, 105
 - of a map, 105
 - surface theory
 - fundamental theorem of, 132
- Special linear affine map, 71
- Special orthogonal affine map, 71
- Sphere, 151
- Standard parameterization, 148
- Steiner, J., 186
- Striction curve, 148, 183
- Structural equation(s)
 - first, 16
 - of SO(3), 71, 73
 - second, 16
- Support
 - function, 184
 - plane, 64

- Surface
 - approximate up to second order, 36
 - classical flat, 141
 - compact of constant negative curvature, 292
 - complete, of constant curvature, 233 ff.
 - developable, 197, 236
 - classical classification of, 237
 - doubly ruled, 153, 155
 - flat, 141
 - minimal, 167
 - of revolution, 156
 - canonical parameterization of, 158
 - geodesics on, 214
 - of constant curvature, 161
 - quadratic, quadric, 118, 151
 - Riemann, 295
 - ruled, 146
 - special affine geometry of, 71
 - triply orthogonal system of, 204
- Surface theory, fundamental theorem of, 56, 73, 74
- Switcheroo, familiar old, 52
- Synge's inequality, 6
- Tangent developable, 142
 - striction curve of, 183
- Tangent to M , 1
- Tangential projection, 1
- Tensor analysis treatment of submanifolds, 12
- Terquem-Joachimsthal theorem, 203
- Theorema Egregium, 55, 69
- Third fundamental form, 62
 - of a map, 62
- Tiling, 296
- Topset, 281
- Torsion, geodesic, 191
- Torus
 - flat, 61
 - g -holed, 292
 - geodesics on, 230
 - umbilics on, 160, 198
- Total
 - absolute curvature, 278
 - curvature, 286
- Totally geodesic, 24
- Tractrix, 164
- Triangulation, 222, 270
- Triply orthogonal system of surfaces, 204
- Tschebyscheff net, 250
 - asymptotic, 251
- Tubular neighborhood, 299
- Type of a point, 78
- Umbilic, 50, 136
 - on ellipsoid, 152
 - lines of curvature in a neighborhood of, 198
 - on elliptic hyperboloid, 154
 - on elliptic paraboloid, 155
 - on torus, 160
 - surface with all points, 51
- Umlaufsatz, *see* Hopf Umlaufsatz
- Uniformization theorem, general, 297
- Unit normal field, 7
- Unknotted, 291
- Vertices, 266
- Voss, K., 290
- Weingarten equations, 7, 8, 9, 14, 20, 53, 122
- Wunderlich, W., 239

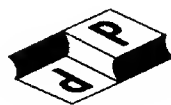
A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME FOUR
Third Edition



MICHAEL SPIVAK

PUBLISH OR PERISH, INC.



Houston, Texas 1999

Publish or Perish, Inc.
www.mathpop.com

Copyright © 1970, 1979, 1999 by Michael Spivak
All Rights Reserved

Volume 1 ISBN 0-914098-70-5
Volume 2 ISBN 0-914098-71-3
Volume 3 ISBN 0-914098-72-1
Volume 4 ISBN 0-914098-73-X
Volume 5 ISBN 0-914098-74-8

Printed in the United States of America

TABLE OF CONTENTS

Although the chapters are not divided into sections,
except for the major subdivisions within Chapter 7,
the listing for each chapter gives some indication
which topics are treated, and on what pages.

CHAPTER 7. HIGHER DIMENSIONS AND CODIMENSIONS

A. THE GEOMETRY OF CONSTANT CURVATURE MANIFOLDS

The standard models of $S^n(K_0)$ and $H^n(K_0)$ in \mathbb{R}^{n+1}	1
Stereographic projection and the conformal model of H^n	5
Conformal maps of \mathbb{R}^n and the isometries of H^n	8
Totally geodesic submanifolds and geodesic spheres of H^n	10
Horospheres and equidistant hypersurfaces	14
Geodesic mappings; the projective model of H^n ; Beltrami's theorem	17

B. CURVES IN A RIEMANNIAN MANIFOLD

Frenet frames and curvatures	21
Curves whose j^{th} curvature vanish	27

C. THE FUNDAMENTAL EQUATIONS FOR SUBMANIFOLDS

The normal connection and the Weingarten equations	32
Second fundamental forms and normal fundamental forms; the Codazzi-Mainardi equations	35
The Ricci equations	38
The fundamental theorem for submanifolds of Euclidean space . . .	41
The fundamental theorem for submanifolds of constant curvature manifolds	50

D. FIRST CONSEQUENCES

The curvatures of a hypersurface; Theorema Egregium; formula for the Gaussian curvature	64
The mean curvature normal; umbilics; all-umbilic submanifolds of Euclidean space	71
All-umbilic submanifolds of constant curvature manifolds	75
Positive curvature and convexity	79

E. FURTHER RESULTS

Flat ruled surfaces in \mathbb{R}^m	85
Flat ruled surfaces in constant curvature manifolds	86
Curves on hypersurfaces	88

F. COMPLETE SURFACES OF CONSTANT CURVATURE

Modifications of results for surfaces in \mathbb{R}^3	91
Surfaces of constant curvature in S^3	93
surfaces with constant curvature 0	94
the Hopf map	107
Surfaces of constant curvature in H^3	110
Jörgens theorem; surfaces of constant curvature 0	112
surfaces of constant curvature -1	117
rotation surfaces of constant curvature between -1 and 0	118

G. HYPERSURFACES OF CONSTANT CURVATURE IN
HIGHER DIMENSIONS

Hypersurfaces of constant curvature in dimensions > 3	119
The Ricci tensor; Einstein spaces, hypersurfaces which are Einstein spaces	120
Hypersurfaces of the same constant curvature as the ambient manifold	123
Addendum 1. The Laplacian	128
Addendum 2. The $*$ operator and the Laplacian on forms; Hodge's Theorem	139
Addendum 3. When are two Riemannian manifolds isometric? . . .	152
Addendum 4. Better imbedding invariants	163
Problems	189

CHAPTER 8. THE SECOND VARIATION

Two-parameter variations; the second variation formula	201
Jacobi fields; conjugate points	208
Minimizing and non-minimizing geodesics	214
The Hadamard-Cartan Theorem	223
The Sturm Comparison Theorem; Bonnet's Theorem	226
Generalizations to higher dimensions; the Morse-Schoenberg Comparison Theorem; Meyer's Theorem; the Rauch Comparison Theorem	230

Synge's lemma; Synge's Theorem	239
Cut points; Klingenberg's theorem	246
Problems	258

CHAPTER 9. VARIATIONS OF LENGTH, AREA, AND VOLUME

Variation of area for normal variations of surfaces in \mathbb{R}^3 ; minimal surfaces	259
Isothermal coordinates on minimal surfaces; Bernstein's Theorem	264
Weierstrass-Enneper representation	268
Associated minimal surfaces; Schwarz's Theorem	274
Change of orientation; Henneberg's minimal surface	276
Classical calculus of variations in n dimensions	281
Variation of volume formula	286
Isoperimetric problems	293
Addendum 1. Isothermal coordinates	314
Addendum 2. Immersed spheres with constant mean curvature	347
Addendum 3. Imbedded surfaces with constant mean curvature	351
Addendum 4. The second variation of volume	355

MINI-BIBLIOGRAPHY FOR VOLUME IV	379
---	-----

NOTATION INDEX	381
--------------------------	-----

INDEX	385
-----------------	-----

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME FOUR

CHAPTER 7

HIGHER DIMENSIONS AND CODIMENSIONS

The aim of this chapter is, roughly speaking, to see whether and how the results of the previous chapters generalize; instead of surfaces in \mathbb{R}^3 , we will be considering higher dimensional manifolds, of higher codimensions, imbedded or immersed in more general Riemannian manifolds. Even at the risk of making the chapter somewhat disorganized, I have tried to make it pretty complete, so that readers do not have to sit gnawing their thumbs wondering whether a generalization does not appear because it is trivial or because it is false, or because it is unknown. It should be mentioned, however, that a few diddly topics, like the Dupin indicatrix, aren't considered at all. In addition, a few points are taken up in later chapters, and the bibliography for appropriate sections should also be consulted. Finally, the most notable omission of all is the generalization of the Gauss-Bonnet Theorem, which occupies the place of honor in the last chapter of the book.

A. THE GEOMETRY OF CONSTANT CURVATURE MANIFOLDS

Although our aim in this chapter is to obtain results of the greatest possible generality, many of the theorems will not hold, or even make sense, unless the ambient manifold has constant curvature K_0 . It will be necessary for us to be as familiar with the properties of these Riemannian manifolds as we are with the case of Euclidean space ($K_0 = 0$). We will consider only the simply-connected complete n -dimensional Riemannian manifolds $(M, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 ; by Problem 1-5, the manifold $(M, \langle \cdot, \cdot \rangle)$ is then uniquely determined up to isometry by K_0 .

For $K_0 > 0$, the manifold $(M, \langle \cdot, \cdot \rangle)$ is just the n -sphere $S^n(K_0)$ of radius $1/\sqrt{K_0}$ in \mathbb{R}^{n+1} ,

$$S^n(K_0) = \left\{ p \in \mathbb{R}^{n+1} : \langle p, p \rangle = \frac{1}{K_0} \right\},$$

with the Riemannian metric induced from the ordinary metric $\langle \cdot, \cdot \rangle$ of \mathbb{R}^{n+1} . For simplicity, we usually consider only the case $K_0 = 1$, setting $S^n = S^n(1)$.

It is clear that every orthogonal map $A \in O(n+1)$ takes S^n to itself and is an isometry. Moreover, $O(n+1)$ is precisely the set of isometries of S^n , since a suitable $A \in O(n+1)$ takes any orthonormal frame $X_1, \dots, X_n \in S^n_p$ at any point $p \in S^n$ to any other orthonormal frame $Y_1, \dots, Y_n \in S^n_q$ at any point $q \in S^n$, and an isometry of S^n is determined by its action on S^n_p (Problem 1-5).

For $K_0 < 0$, we can obtain an analogous submanifold of \mathbb{R}^{n+1} by considering a non-positive definite Riemannian metric on \mathbb{R}^{n+1} . Denoting the components of a point $a \in \mathbb{R}^{n+1}$ by a^0, a^1, \dots, a^n , we consider first the non-degenerate inner product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^{n+1} defined by

$$\langle a, b \rangle = -a^0 b^0 + a^1 b^1 + \dots + a^n b^n.$$

This is called the **Lorentzian** inner product on \mathbb{R}^{n+1} , and the group $O^1(n+1)$ of all linear transformations $f: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ which preserve $\langle \cdot, \cdot \rangle$ is called the **Lorentz group** [actually (Problem 1), any map $f: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ preserving $\langle \cdot, \cdot \rangle$ is automatically linear]. By means of the standard identification of \mathbb{R}^{n+1}_p with \mathbb{R}^{n+1} , we obtain a non-degenerate inner product $\langle \cdot, \cdot \rangle_p$ on each \mathbb{R}^{n+1}_p , and thus a non-positive definite Riemannian metric on \mathbb{R}^{n+1} , which we denote also simply by $\langle \cdot, \cdot \rangle$. In terms of the standard coordinate system x^0, x^1, \dots, x^n on \mathbb{R}^{n+1} we have

$$\langle \cdot, \cdot \rangle = -dx^0 \otimes dx^0 + dx^1 \otimes dx^1 + \dots + dx^n \otimes dx^n.$$

The isometries of $(\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle)$ are (Problem 2) precisely the maps of the form

$$p \mapsto A(p) + q \quad A \in O^1(n+1), \quad q \in \mathbb{R}^{n+1}.$$

Now for $K_0 < 0$ consider the quadric hypersurface

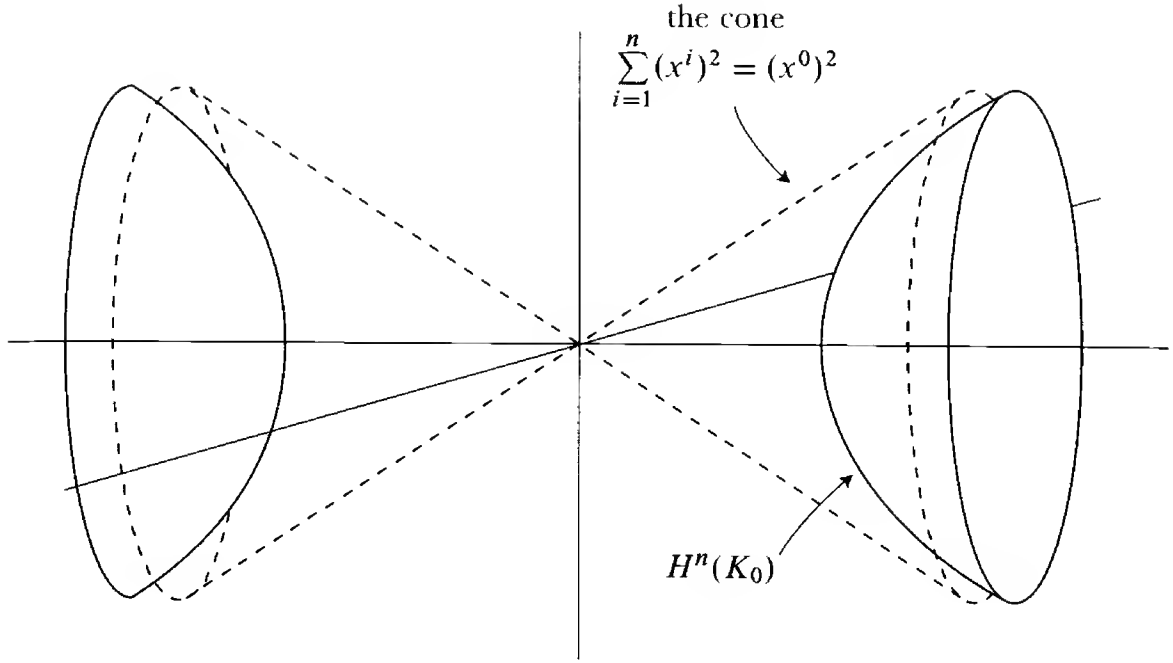
$$\left\{ p \in \mathbb{R}^{n+1} : \langle p, p \rangle = \frac{1}{K_0} \right\}.$$

As illustrated on the next page, this consists of two components, each homeomorphic to \mathbb{R}^n ; we will pick one of them, say the one consisting of points with $p^0 > 0$, and define

$$H^n(K_0) = \left\{ p \in \mathbb{R}^{n+1} : p^0 > 0 \text{ and } \langle p, p \rangle = \frac{1}{K_0} \right\}.$$

For simplicity, we usually consider only the case $K_0 = -1$, setting $H^n(-1) = H^n$, the “ n -dimensional hyperbolic space”. To find the tangent space H^n_p , we proceed precisely as in the case of S^n . Any curve c in H^n satisfies

$$\langle c(t), c(t) \rangle = 0 \text{ for all } t \implies \langle c'(t), c(t) \rangle = 0,$$



so H_p^n contains only vectors v_p with $\langle v, p \rangle = 0$. Moreover, $\{v : \langle v, p \rangle = 0\}$ is the kernel of the non-zero linear functional $v \mapsto \langle v, p \rangle$, so it has dimension exactly $n - 1$. Thus

$$H_p^n = \{v_p : \langle v, p \rangle = 0\} \quad \text{for } p \in H^n, \quad \text{i.e., } \langle p, p \rangle = -1.$$

We now claim that the induced Riemannian metric on H^n is positive definite. To show this, it is convenient to consider the **index** of a bilinear function $B: V \times V \rightarrow \mathbb{R}$ on a vector space V , which is defined to be the largest dimension of any subspace $W \subset V$ on which B is negative definite [that is, $B(w, w) < 0$ for all $0 \neq w \in W$]. The bilinear function

$$(a, b) \mapsto \langle a, b \rangle = -a^0 b^0 + a^1 b^1 + \cdots + a^n b^n$$

on \mathbb{R}^{n+1} clearly has index ≥ 1 , for it is negative definite on the subspace $U^- = \{(a^0, 0, \dots, 0)\}$. Moreover, $\langle \cdot, \cdot \rangle$ is positive definite on the subspace $U^+ = \{(0, a^1, \dots, a^n)\}$. If $\langle \cdot, \cdot \rangle$ were negative definite on a subspace W of dimension ≥ 2 , then $\langle \cdot, \cdot \rangle$ would be negative definite on the non-zero subspace $W \cap U^+$, which is clearly impossible. So $\langle \cdot, \cdot \rangle$ has index 1. Naturally, each $\langle \cdot, \cdot \rangle_p$ also has index 1. Now consider $\langle \cdot, \cdot \rangle_p$ on H_p^n . If $v_p \in H_p^n$, then v is linearly independent of p , and we already have $\langle p, p \rangle < 0$, so we cannot have $\langle v, v \rangle < 0$, as $\langle \cdot, \cdot \rangle$ has index 1. Nor can we even have $\langle v, v \rangle = 0$, for then we would have

$$\langle p + v, p + v \rangle = \langle p, p \rangle + 2\langle p, v \rangle + \langle v, v \rangle = \langle p, p \rangle < 0,$$

which is also impossible. Thus $\langle \cdot, \cdot \rangle_p$ is positive definite on H^n_p , and H^n is an ordinary Riemannian manifold. (In the picture on the previous page this is quite clear, since all tangent lines have greater slope than the generators of the cone $\sum_i (x^i)^2 = (x^0)^2$, and a vector v along one of these generators satisfies $\langle v, v \rangle = 0$.)

Naturally every element of $O^1(n+1)$ which keeps $\{p \in \mathbb{R}^{n+1} : p^0 > 0\}$ fixed will give an isometry of H^n onto itself. We also claim that all isometries of H^n arise in this way. To prove this, we just note that if $(v_1)_p, \dots, (v_n)_p \in H^n_p$ is orthonormal, and similarly for $(w_1)_q, \dots, (w_n)_q \in H^n_q$, so that

$$\begin{aligned}\langle p, p \rangle &= \langle q, q \rangle = -1 \\ \langle v_i, p \rangle &= 0 = \langle w_i, q \rangle \\ \langle v_i, v_j \rangle &= \langle w_i, w_j \rangle = \delta_{ij},\end{aligned}$$

then the linear transformation taking

$$p \mapsto q \quad \text{and} \quad v_i \mapsto w_i$$

is clearly in $O^1(n+1)$. Since there are thus isometries of H^n taking any orthonormal basis at any point to any orthonormal basis at any other point, H^n must have constant curvature. We can compute that $H^n(K_0)$ has constant curvature K_0 in a manner exactly analogous to a computation of the curvature of $S^n(K_0)$, by using Theorems 1-1, 1-6, and 1-9; the only difference is that we must allow the ambient manifold in Theorems 1-1 and 1-6 to have a non-positive definite Riemannian metric, and the “unit” normal field v in 1-9 will actually satisfy $\langle v, v \rangle = -1$. The manifold $H^n(K_0)$ is (geodesically) complete. Because we are dealing with an indefinite metric on \mathbb{R}^{n+1} , this does not simply follow from the fact that $H^n(K_0)$ is a closed subset of \mathbb{R}^{n+1} . However, it is an easy exercise to prove completeness using the fact that there are isometries taking any orthonormal basis to any other. We also mention that the geodesics of H^n are (Problem 3) precisely the intersections $H^n \cap P$ where P is a plane in \mathbb{R}^{n+1} through 0; more generally, the totally geodesic submanifolds of H^n are $H^n \cap P$ where P is a vector subspace of \mathbb{R}^{n+1} .

In the past we have given several other models for H^n , and for S^n minus a point. For example, we have described the metric of a space of constant curvature K_0 in terms of normal coordinates, in Addendum 1 to Chapter II.7. In Addendum 2 to that chapter we found the most general isothermal coordinate systems on the manifolds of constant curvature, after first determining the expression for the metric on S^n in the coordinate system defined by “stereographic projection”. To define this map, we considered S^n as the sphere of radius 1

around the point $(0, \dots, 0, 1)$, so that S^n is tangent to $\mathbb{R}^n = \mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+1}$. Letting $*$ be the “north pole” $* = (0, \dots, 0, 2) \in S^n$, the **stereographic projection**

$$\sigma: S^n - \{*\} \rightarrow \mathbb{R}^n$$

is defined geometrically as follows: for any $p \neq *$ in S^n , we let $\sigma(p)$ be the point where the line between p and $*$ intersects \mathbb{R}^n . It is easy to check (see the

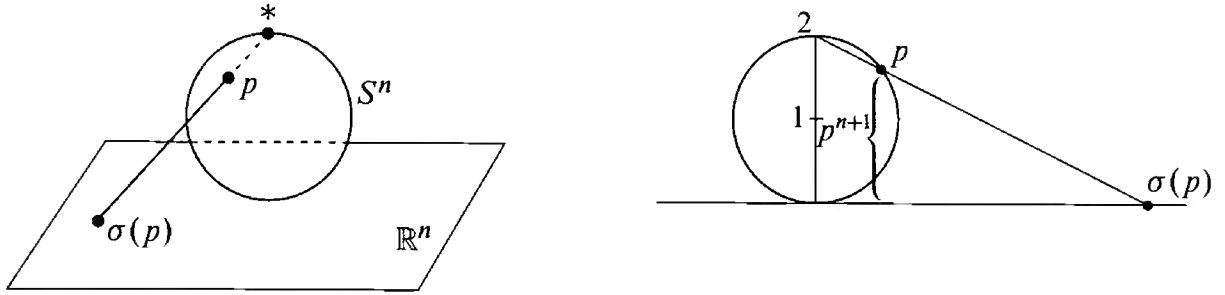


figure on the right) that

$$(1) \quad \sigma(p) = \left(\frac{2p^1}{2 - p^{n+1}}, \dots, \frac{2p^n}{2 - p^{n+1}} \right)$$

and that $f = \sigma^{-1}$ is given by

$$(2) \quad \sigma^{-1}(y) = f(y) = \left(\frac{y^1}{1 + \frac{1}{4} \sum_i (y^i)^2}, \dots, \frac{y^n}{1 + \frac{1}{4} \sum_i (y^i)^2}, \frac{\frac{1}{2} \sum_i (y^i)^2}{1 + \frac{1}{4} \sum_i (y^i)^2} \right).$$

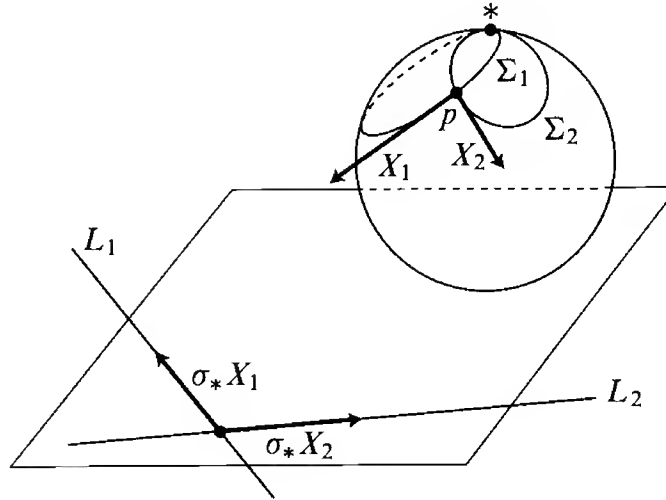
If y^1, \dots, y^n denotes the standard coordinate system on \mathbb{R}^n , then the $y^i \circ \sigma$ give a coordinate system on $S^n - \{*\}$. We can compute the metric $\langle \cdot, \cdot \rangle$ in terms of this coordinate system by computing

$$\begin{aligned} f^* \sum_{i=1}^{n+1} dx^i \otimes dx^i &= \sum_{i=1}^{n+1} df^i \otimes df^i \\ &= \sum_{i=1}^{n+1} \sum_{j,k=1}^n \frac{\partial f^i}{\partial y^j} \frac{\partial f^i}{\partial y^k} dy^j \otimes dy^k, \end{aligned}$$

by means of equation (2).

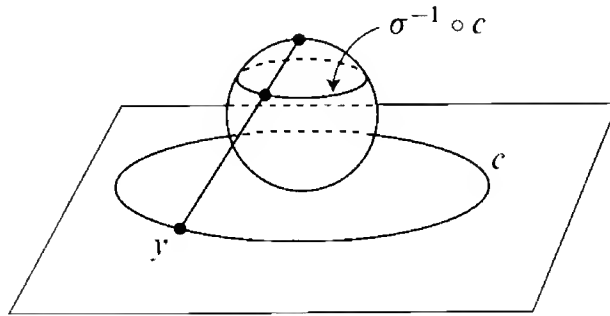
However we can save ourselves a lot of computational work by first proving geometrically that σ is conformal. It clearly suffices to consider the case $n = 2$.

Notice first that if $L \subset \mathbb{R}^2$ is a straight line, then the lines through $*$ and points of L form a plane with a horizontal line through $*$ deleted, so $\sigma^{-1}(L)$ is $\Sigma - \{*\}$ for some circle $\Sigma \subset S^2$. Now given two linearly independent vectors $X_1, X_2 \in S^2_p$, consider the straight lines L_1, L_2 through $\sigma(p)$ pointing in the directions of $\sigma_*(X_1)$ and $\sigma_*(X_2)$. Their inverse images under σ are $\Sigma_1 - \{*\}$



and $\Sigma_2 - \{*\}$ for two circles $\Sigma_1, \Sigma_2 \subset S^2$ containing $*$. The angle between X_1 and X_2 is the angle of intersection of Σ_1 and Σ_2 at p , which is the same as the angle of intersection of Σ_1 and Σ_2 at $*$. But the tangent lines to Σ_1 and Σ_2 at $*$ are parallel to L_1 and L_2 , respectively. So the angle of intersection at $*$ is the same as the angle between σ_*X_1 and σ_*X_2 . Thus, σ is conformal.

Now for any point $y \in \mathbb{R}^n$, let $c: [0, 2\pi] \rightarrow \mathbb{R}^n$ be a curve, parameterized proportionally to arclength, which goes once around a circle centered at 0 and passing through y ; thus c' always has squared length $|y|^2$. Formula (2) shows



that $(\sigma^{-1} \circ c)'$ always has squared length

$$\sum_{i=1}^n \left[\frac{y^i}{1 + \frac{1}{4} \sum_i (y^i)^2} \right]^2 = \frac{|y|^2}{\left[1 + \frac{1}{4} \sum_i (y^i)^2 \right]^2}.$$

This shows that in the conformal coordinate system $\{x^i = y^i \circ \sigma\}$ on $S^n - \{*\}$, the metric $\langle \cdot, \cdot \rangle$ has the form

$$(3) \quad \langle \cdot, \cdot \rangle = \sum_{i=1}^n \frac{dx^i \otimes dx^i}{\left[1 + \frac{1}{4} \sum_i (x^i)^2\right]^2}.$$

If we were dealing with a sphere of curvature K_0 , the factor $1/4$ would be replaced by $K_0/4$.

For later use, we mention one further property of the stereographic projection: it takes spheres in S^n to spheres and hyperplanes of \mathbb{R}^n , and *vice-versa*. Indeed, a sphere $\Sigma \subset S^n$ is the intersection of S^n with some hyperplane,

$$\Sigma = \left\{ p \in S^n : \sum_{i=1}^{n+1} \alpha_i p^i = \beta \right\},$$

and then

$$\begin{aligned} y \in \sigma(\Sigma) &\iff \sigma^{-1}(y) \in \Sigma \\ &\iff \sum_{i=1}^n \alpha_i \frac{y^i}{1 + \frac{1}{4} \sum_i (y^i)^2} + \frac{1}{2} \alpha_{n+1} \frac{\sum_i (y^i)^2}{1 + \frac{1}{4} \sum_i (y^i)^2} = \beta, \quad \text{by (2).} \end{aligned}$$

This is always a sphere or hyperplane in \mathbb{R}^n , and the converse works similarly.

Now for $K_0 < 0$, in particular for $K_0 = -1$, we can just formally replace the factor $1/4$ in (3) by $-1/4$. In Addendum 2 to Chapter II.7 we showed that this metric does indeed have $K_0 = -1$. In fact, this metric was simply one possible choice for the conformal metrics of constant curvature $K_0 = -1$.

We have already pointed out that, in order to have a connected manifold, we must consider the metric

$$\langle \cdot, \cdot \rangle = \sum_{i=1}^n \frac{dx^i \otimes dx^i}{\left[1 - \frac{1}{4} \sum_i (x^i)^2\right]^2}$$

only on the open ball of radius 2,

$$B^n = B^n(2) = \{x \in \mathbb{R}^n : \sum_i (x^i)^2 < 4\},$$

but that $\langle \cdot, \cdot \rangle$ is already complete on B^n (see pg. II.339). Thus $(B^n, \langle \cdot, \cdot \rangle)$ must be isometric to the space $H^n \subset (\mathbb{R}^{n+1}, \langle \cdot, \cdot \rangle)$; a method for constructing an explicit isometry between $(B^n, \langle \cdot, \cdot \rangle)$ and H^n will be suggested later.

The model $(B^n, \langle \cdot, \cdot \rangle)$ will often be very useful, and we will examine it in great detail, determining, in particular, precisely what the isometries of $(B^n, \langle \cdot, \cdot \rangle)$ onto itself look like. In order to do this, however, we first need to generalize a few results from previous chapters.

First of all, Dupin's Theorem (4-10) on triply orthogonal systems of surfaces generalizes immediately to a theorem on n -orthogonal systems of hypersurfaces in \mathbb{R}^n . We will also need to generalize Theorem 2-2, concerning all-umbilic surfaces in \mathbb{R}^3 . For a hypersurface $M \subset \mathbb{R}^{n+1}$ we locally have a unit normal field $\nu: M \rightarrow S^n \subset \mathbb{R}^{n+1}$, and a map $d\nu: M_p \rightarrow M_p$ (Theorem 1-8); we call $p \in M$ an **umbilic** if $d\nu: M_p \rightarrow M_p$ is multiplication by a constant.

1. LEMMA. For $n \geq 2$, let $M \subset \mathbb{R}^{n+1}$ be a connected hypersurface with all points umbilics. Then M is part of a hyperplane or an n -dimensional sphere.

Remark: Later on we will have much more general results.

PROOF. As in the proof of Theorem 2-2, it suffices to prove this locally. Choose an adapted orthonormal moving frame $X_1, \dots, X_n, X_{n+1} = \nu$ on M . By hypothesis, there is a function λ on M such that

$$(1) \quad \nabla'_X X_{n+1} = -\lambda X \quad X \text{ tangent to } M.$$

In terms of the dual and connection forms we have

$$\psi_{n+1}^j(X) = \langle \nabla'_X X_{n+1}, X_j \rangle = -\lambda \langle X, X_j \rangle,$$

and thus

$$\psi_j^{n+1} = -\psi_{n+1}^j = \lambda \theta^j.$$

Taking the exterior derivative of this equation, we obtain

$$\begin{aligned} d\lambda \wedge \theta^j + \lambda d\theta^j &= d\psi_j^{n+1} = -\sum_i \psi_i^{n+1} \wedge \omega_j^i \quad (\text{pg. III.19}) \\ &= -\lambda \sum_i \theta^i \wedge \omega_j^i, \end{aligned}$$

while

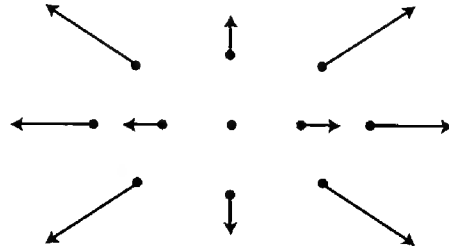
$$d\theta^j = -\sum_i \omega_j^i \wedge \theta^i.$$

So we find that

$$d\lambda \wedge \theta^j = 0 \quad j = 1, \dots, n.$$

This implies that $d\lambda = 0$, so λ is constant.

The remainder of the argument can be carried out as in the proof of Theorem 2-2, by considering an immersion $f: U \rightarrow M$ for $U \subset \mathbb{R}^n$ open. Here is an alternative (essentially equivalent) argument. If $\lambda = 0$, then all $\psi_j^{n+1} = 0$, so the second fundamental form $s = 0$; thus M is totally geodesic (Propositions 1-16 and 1-17), so M lies in a hyperplane. So we assume $\lambda \neq 0$. Let V be the vector field on \mathbb{R}^{n+1} defined by

$$V(p) = p_p \in \mathbb{R}^{n+1}_p.$$


If x^1, \dots, x^{n+1} is the standard coordinate system on \mathbb{R}^{n+1} , then

$$V = \sum_i x^i \frac{\partial}{\partial x^i},$$

and we easily see that $\nabla'_X V = X$ for all tangent vectors X of \mathbb{R}^{n+1} . Thus equation (1) can be written

$$\nabla'_X (X_{n+1} + \lambda V) = 0.$$

Thus the vector field $X_{n+1} + \lambda V$ is parallel along M . Identifying tangent vectors of \mathbb{R}^{n+1} with elements of \mathbb{R}^{n+1} , this means that $X_{n+1} + \lambda V$ is a constant vector v_0 on M , so we have

$$X_{n+1}(p) + \lambda p = v_0 \in \mathbb{R}^{n+1}.$$

Thus

$$p = \frac{v_0 - X_{n+1}(p)}{\lambda}$$

for all $p \in M$, which means that M lies in the sphere of radius $1/\lambda$ around the point v_0/λ . ♦

Using Lemma 1. and the generalization of Dupin's Theorem, it is now a straightforward matter to generalize Liouville's Theorem (4-12) to \mathbb{R}^n : every conformal map of an open subset of \mathbb{R}^n onto an open subset of \mathbb{R}^n is the restriction of a composition of similarities and inversions, in fact at most one of each. In addition (compare the proof of Lemma 4-13), these conformal maps take hyperplanes and spheres to hyperplanes and spheres.

With this information we are now in a good position to consider the isometries of $(B^n, \langle \cdot, \cdot \rangle)$. Since $\langle \cdot, \cdot \rangle$ is conformally equivalent to the usual metric $\sum_i dx^i \otimes dx^i$ on B^n , we see immediately that

- (1) Every isometry $f: (B^n, \langle \cdot, \cdot \rangle) \rightarrow (B^n, \langle \cdot, \cdot \rangle)$ onto itself is a conformal map of B^n onto itself (as a subset of \mathbb{R}^n with the usual metric).

We next claim

- (2) If $f: B^n \rightarrow B^n$ is a conformal map of B^n onto itself and $f_*: B^n_p \rightarrow B^n_p$ is a multiple of the identity for some $p \in B^n$, then f is the identity (or possibly minus the identity, if $p = 0$).

To prove this, consider the sphere $S = \text{boundary } B^n$. Then S is taken into itself by f (more precisely, by the composition of similarities and inversions of which f is the restriction). If P is a hyperplane through p , then $f(P)$ is a hyperplane or sphere tangent to P at p (since $f_*: B^n_p \rightarrow B^n_p$ is a multiple of the identity). But also the angle at which P cuts S must equal the angle at which $f(P)$ cuts $f(S) = S$. It follows easily that $f(P) = P$. Consequently, f cannot be an inversion or the composition of one inversion and one similarity, for the inversion must be through a point $p_* \notin B^n$, and then $f(P)$ could not be a plane. So f must be a similarity, and the desired result follows easily.

Now consider any conformal map $f: B^n \rightarrow B^n$ of B^n onto itself, and let $p \in B^n$ be a point $\neq 0$. If $X_1, \dots, X_n \in B^n_p$ is an orthonormal basis with respect to $\langle \cdot, \cdot \rangle_p$, then there is some $\lambda > 0$ with

$$\langle f_*(X_i), f_*(X_j) \rangle_{f(p)} = \lambda \cdot \delta_{ij},$$

so $\{f_*(X_i)/\sqrt{\lambda}\}$ is an orthonormal basis for $B^n_{f(p)}$. Consequently, there is an isometry $g: (B^n, \langle \cdot, \cdot \rangle) \rightarrow (B^n, \langle \cdot, \cdot \rangle)$ with

$$g_*(X_i) = \frac{1}{\sqrt{\lambda}} f_*(X_i).$$

Then $g^{-1} \circ f: B^n \rightarrow B^n$ is a conformal map of B^n onto itself (by (1)), and $(g^{-1} \circ f)_*: B^n_p \rightarrow B^n_p$ is a multiple of the identity. So $g = f$ by (2). Thus

- (3) Every conformal map $f: B^n \rightarrow B^n$ of B^n onto itself is an isometry of $(B^n, \langle \cdot, \cdot \rangle)$ onto itself.

We can now deduce some further information about $(B^n, \langle \cdot, \cdot \rangle)$. We know (pg. III.26) that the d -dimensional totally geodesic submanifolds through $0 \in B^n$ are just $B^n \cap P$, where P is a d -dimensional plane through 0 in \mathbb{R}^n . Now

any totally geodesic submanifold is the image of $B^n \cap P$ under some isometry $f: B^n \rightarrow B^n$. The map f is conformal by (1), so

- (4) Every totally geodesic submanifold of B^n is the intersection of B^n with a plane or sphere which intersects $S = \text{boundary } B^n$ orthogonally.

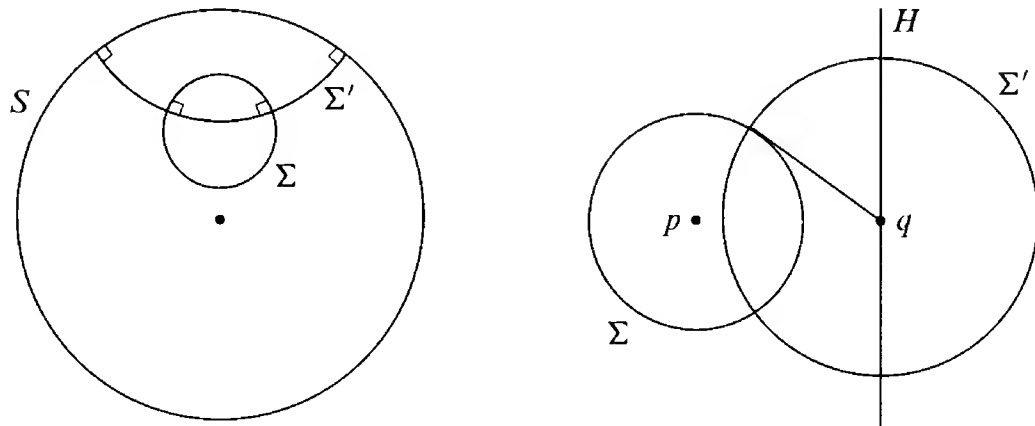
Conversely, suppose that Σ is a plane or sphere which intersects S orthogonally, and let $p \in \Sigma \cap B^n$. There is a totally geodesic submanifold of B^n tangent to Σ at p . By (4), this submanifold must intersect S orthogonally. So it must be precisely $\Sigma \cap B^n$. Thus

- (5) The intersection with B^n of a plane or sphere which intersects S orthogonally is a totally geodesic submanifold.

Next consider a geodesic sphere Σ around $0 \in B^n$ (that is, let Σ be the set of points at fixed $\langle \cdot, \cdot \rangle$ distance from 0). By symmetry of $\langle \cdot, \cdot \rangle$, the set Σ is an ordinary (hyper) sphere. Now any geodesic sphere is the image of Σ under some isometry $f: B^n \rightarrow B^n$. Since this isometry is a conformal map we see that

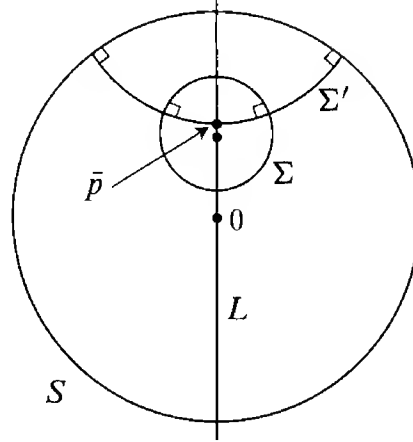
- (6) Every geodesic sphere of $(B^n, \langle \cdot, \cdot \rangle)$ is an ordinary hypersphere completely contained in B^n .

Now we will work on proving the converse of (6). Suppose we have an ordinary hypersphere Σ completely contained in B^n . We claim first of all that there is a hypersphere Σ' which is orthogonal to both Σ and S . To prove this



we note that by means of an inversion through a point of S , we can reduce the problem to that of finding a hypersphere Σ' orthogonal to a hyperplane H and a hypersphere Σ lying completely on one side of it. If Σ has center p and $q \in H$ is the point closest to p , then we simply choose Σ' to be a hypersphere around q whose radius has the length of a tangent from q to Σ . Now

that we have the hypersphere Σ' orthogonal to both Σ and S , we consider the intersection \bar{p} of Σ' and the line L between 0 and the center of Σ . Let



$f: B^n \rightarrow B^n$ be an isometry taking \bar{p} to 0. We know by (5) that $B^n \cap \Sigma'$ is a totally geodesic hypersurface. Therefore f must take Σ' to a hyperplane H through 0. Moreover, f takes the geodesic L to another geodesic through 0, i.e., to a straight line L' through 0, but not lying in H . The image hypersphere $f(\Sigma)$ must be perpendicular to both H and L' , which can happen only when $f(\Sigma)$ has center 0. So $f(\Sigma)$ is a geodesic sphere, which implies that Σ is also:

- (7) Every ordinary hypersphere completely contained in B^n is a geodesic sphere.

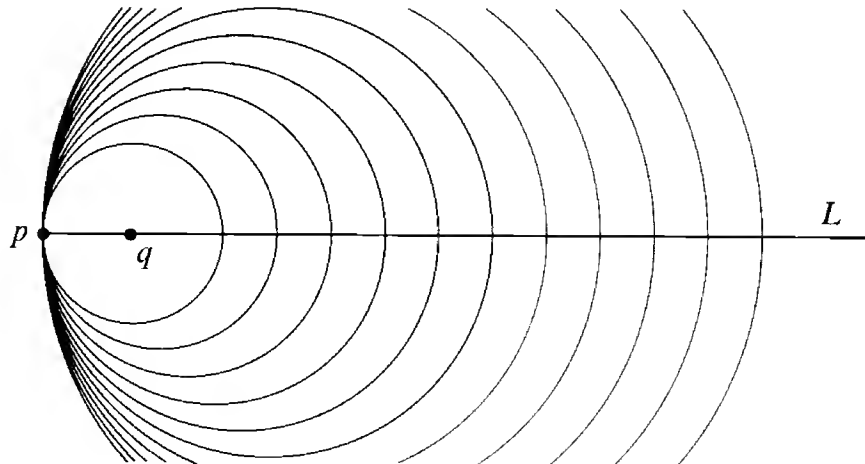
All of this information, by the way, was obtained only for the case $n \geq 3$, since we made use of Liouville's Theorem. The case $n = 2$ is sometimes analyzed by explicit computation, making use of the identification of \mathbb{R}^2 with \mathbb{C} (see Problems 4, 5, 6), but we can also use the information which we already have for $n \geq 3$. To do this we consider B^2 as a totally geodesic surface in B^3 . An isometry $f: B^2 \rightarrow B^2$ of B^2 onto itself clearly extends to an isometry $\tilde{f}: B^3 \rightarrow B^3$ of B^3 onto itself. Since \tilde{f} is conformal, f is also. Moreover, since \tilde{f} is a composition of at most one similarity and inversion, it is not hard to see that the same must be true of f (this information is not redundant in the 2-dimensional case). Conversely, if $f: B^2 \rightarrow B^2$ is a conformal map of B^2 onto itself which happens to be a composition of at most one similarity and inversion, then f can easily be extended to a similar conformal map $\tilde{f}: B^3 \rightarrow B^3$ of B^3 onto itself. Since \tilde{f} is an isometry, so is f . It now follows, exactly as before, that the geodesics of B^2 are portions of lines or circles intersecting S orthogonally, while the geodesic circles are the ordinary circles completely contained in B^2 .

In Addendum 2 to Chapter II.7 we also described a complete manifold of constant curvature $K_0 = -1$ by means of the metric

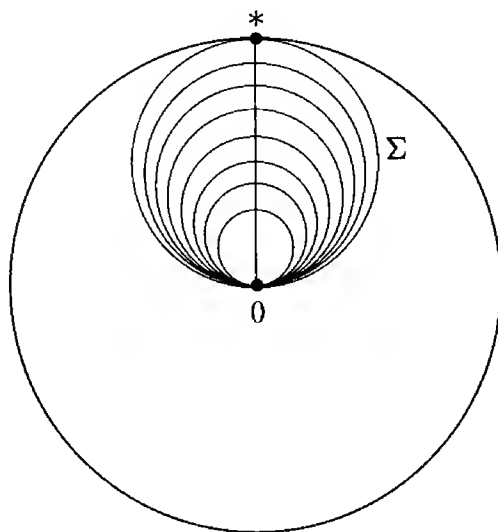
$$\sum_{i=1}^n \frac{dx^i \otimes dx^i}{(x^n)^2}$$

on the upper half-space $\mathcal{H}^n = \{a \in \mathbb{R}^n : a^n > 0\}$. It is easy to describe the isometry between \mathcal{H}^n and B^n . In fact, since the metric on each of them is conformally equivalent to the usual metric on \mathbb{R}^n , the isometry $f: B^n \rightarrow \mathcal{H}^n$ must be a conformal map. If we take an inversion I about a point $*$ on the boundary sphere S of B , then $I(B)$ will be an open half-space, and it is only necessary to compose I with an appropriate similarity. We now easily see that the isometries of \mathcal{H}^n onto itself are precisely the conformal maps taking \mathcal{H}^n onto itself; that the totally geodesic submanifolds of \mathcal{H}^n are $\Sigma \cap \mathcal{H}^n$ for planes and spheres Σ intersecting \mathbb{R}^{n-1} orthogonally; and that the geodesic spheres of \mathcal{H}^n are the ordinary spheres completely contained in \mathcal{H}^n . It will prove extremely useful to be able to shuttle back and forth between B^n and \mathcal{H}^n .

We have now given intrinsic characterizations of the sets $\Sigma \cap B^n$ [or $\Sigma \cap \mathcal{H}^n$] when Σ is a hyperplane or hypersphere intersecting S [or \mathbb{R}^{n-1}] either orthogonally, or else not at all. We also want to give intrinsic characterizations when Σ intersects non-orthogonally. There are two different cases to consider, the first of which is related to a certain limiting construction which played an essential role in the earliest investigations of non-Euclidean geometry. Take a ray L , with initial point p , in a non-Euclidean space. For each q on L , consider the sphere with center q that passes through p . As $q \rightarrow \infty$, this sphere approaches a surface. In the Euclidean case, this surface is just the



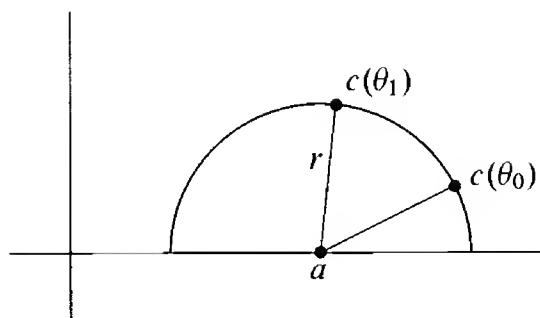
plane through p perpendicular to L ; in the non-Euclidean case, the limiting set is called a “limit sphere” or **horosphere**. It is easy to see that the horospheres of $(B^n, \langle \cdot, \cdot \rangle)$ are precisely $B^n \cap \Sigma$ where Σ is a hypersphere completely inside B except for one point. (First consider the horospheres determined by a ray starting at 0, as in the figure below, and then note that there are isometries of B^n



taking any horosphere to any other.) The early non-Euclidean geometers had their minds blown when they proved that the laws of *Euclidean* geometry hold on the horosphere; in other words, the horosphere is flat. The easiest way for us to see this is to consider an isometry $f: B^n \rightarrow \mathcal{H}^n$ which involves an inversion around the unique point $* \in \Sigma \cap S$. The image $f(\Sigma)$ is then a hyperplane parallel to \mathbb{R}^{n-1} . But the metric induced on this hyperplane is a constant multiple of $\sum_i dx^i \otimes dx^i$, so this hyperplane (which is a horosphere of \mathcal{H}^n) is flat; all other horospheres are isometric images of this one, so they are also flat.

To describe the other sets $\Sigma \cap B^n$ and $\Sigma \cap \mathcal{H}^n$, we first do a short computation in \mathcal{H}^2 . Consider a semi-circle intersecting \mathbb{R}^1 orthogonally, parameterized by

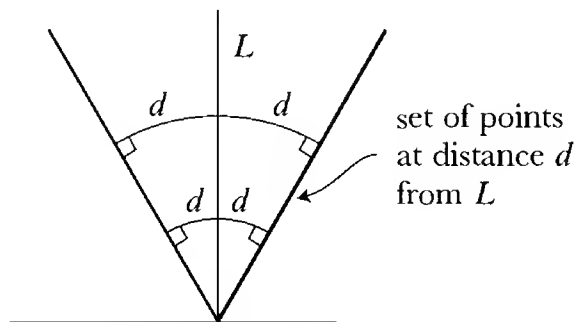
$$c(\theta) = (a + r \cos \theta, r \sin \theta).$$



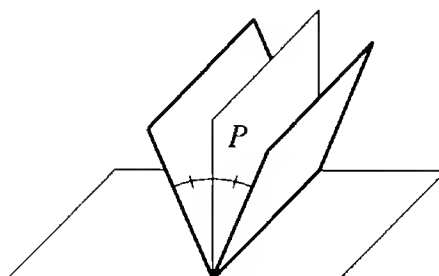
This curve is a geodesic, apart from its parameterization. Its length from the point $c(\theta_0)$ to $c(\theta_1)$ is

$$\begin{aligned} \int_{\theta_0}^{\theta_1} |c'(\theta)| d\theta &= \int_{\theta_0}^{\theta_1} \left| (-r \sin \theta) \frac{\partial}{\partial x^1} + (r \cos \theta) \frac{\partial}{\partial x^2} \right| d\theta \\ &= \int_{\theta_0}^{\theta_1} \sqrt{\frac{(-r \sin \theta)^2 + (r \cos \theta)^2}{(r \cos \theta)^2}} d\theta \\ &= \int_{\theta_0}^{\theta_1} \frac{1}{\cos \theta} d\theta. \end{aligned}$$

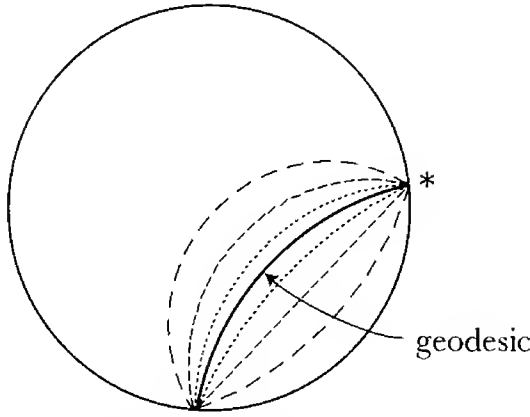
Notice that this is independent of r . It follows that for a geodesic L which is a straight line perpendicular to \mathbb{R} , the set of points at a fixed distance d from L is a pair of straight lines making equal angles with L . Similarly, if P is a totally



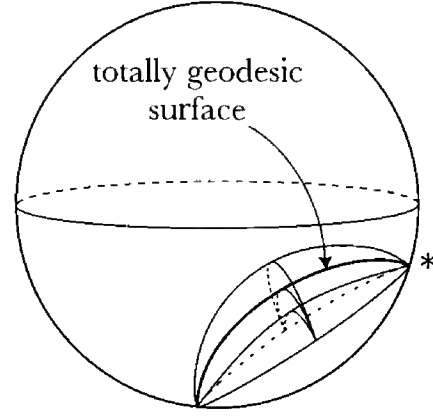
geodesic hypersurface in \mathcal{H}^n consisting of a hyperplane to \mathbb{R}^{n-1} , then the set of points at a fixed distance d from P is a pair of hyperplanes P_1, P_2 making equal angles with P . For the isometry $f: B^n \rightarrow \mathcal{H}^n$, involving an inversion about the



point $* \in S$, the set $f^{-1}(P)$ is $\Sigma \cap B^n$ for some hyperplane or hypersphere Σ with $* \in \Sigma \cap S$; and the sets $f^{-1}(P_i)$ are sets of the same sort. We thus see that



three pairs of lines at fixed distances
from a given geodesic



a pair of surfaces at fixed distance from
a given totally geodesic surface

for hyperplanes or hyperspheres Σ which intersect S [or \mathbb{R}^{n-1}] in more than one point, but not orthogonally, the set $\Sigma \cap B^n$ [or $\Sigma \cap \mathcal{H}^n$] is one component of the set of points at a fixed distance from a totally geodesic hypersurface; these sets are thus called **equidistant hypersurfaces**.

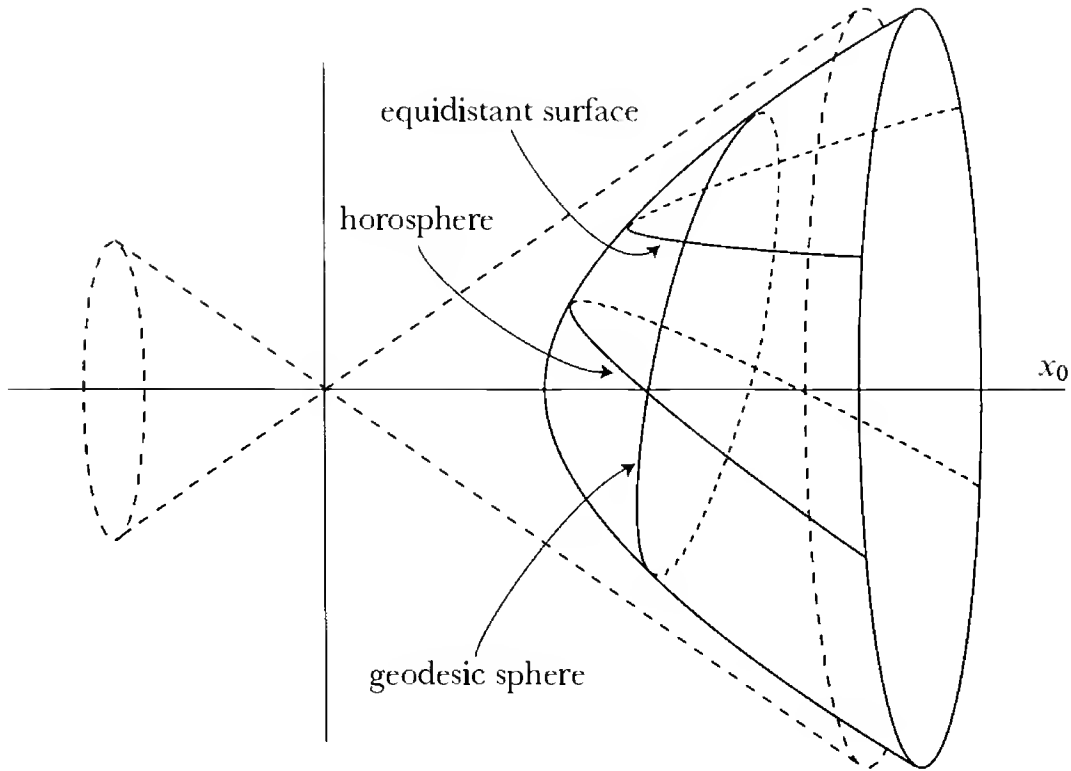
By the way, we can also describe the geodesic spheres, horospheres, and equidistant hypersurfaces for

$$H^n = \{p \in \mathbb{R}^{n+1} : p^0 > 0 \text{ and } \langle p, p \rangle = -1\}.$$

They are all of the form $H^n \cap P$ for some hyperplane P . As illustrated in the figure on the top of the next page, the parabolas, which occur when P is parallel to generator of the cone $\sum_i (x^i)^2 = (x^0)^2$, are horospheres; ellipses, which occur when P makes a larger angle with the x^0 -axis, are geodesic spheres; and hyperbolas, which occur when P makes a smaller angle, are equidistant hypersurfaces. Although these assertions should look pictorially reasonable, we are not yet in a position to prove them (see page 78). For the moment we merely want to note that we would not obtain any new hypersurfaces by considering the sets $H^n \cap Q$ where Q is another quadric hypersurface of the form

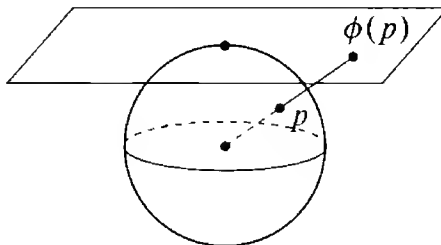
$$Q = \{p \in \mathbb{R}^{n+1} : \langle p - p_0, p - p_0 \rangle = c\};$$

for it is easy to see that $H^n \cap Q$ is always of the form $H^n \cap P$ for some hyperplane P . (Analogously, the intersection of two ordinary spheres in \mathbb{R}^{n+1} is also the intersection of one sphere with a hyperplane.)



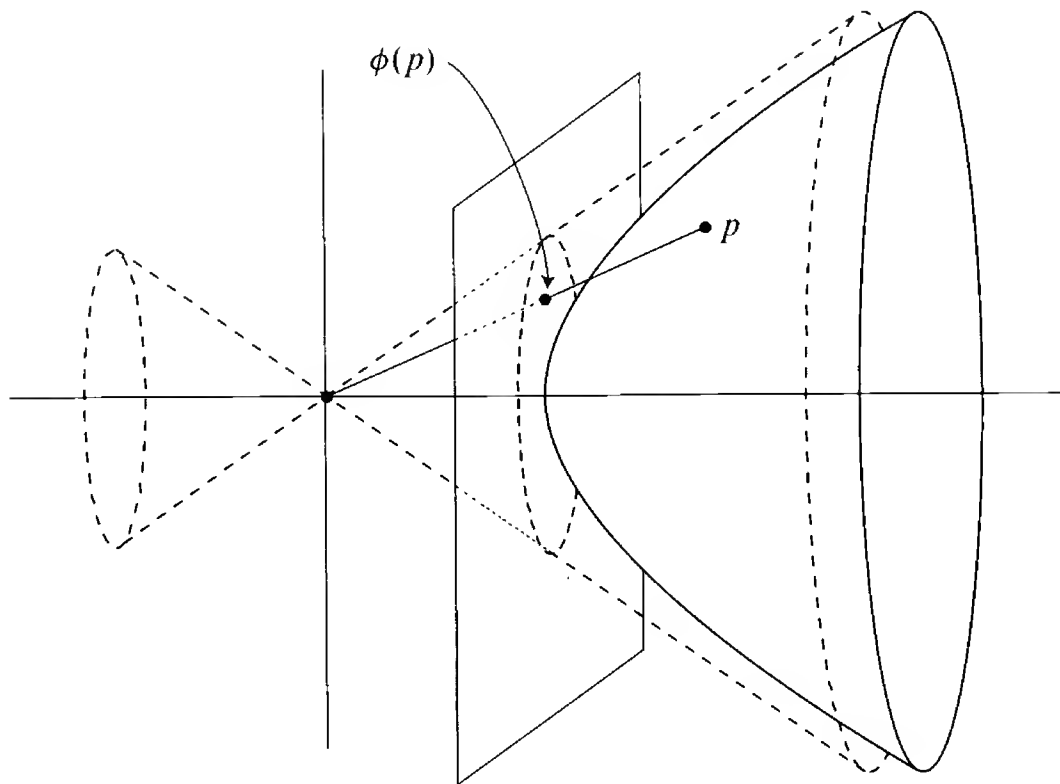
Much of the above discussion evolved from the existence of conformal maps from S^n or H^n to \mathbb{R}^n . Another kind of map will play an important role. A homeomorphism $\phi: M_1 \rightarrow M_2$ from M_1 into M_2 is called a **geodesic mapping** if for every geodesic γ of M_1 , the composition $\phi \circ \gamma$ is a reparameterization of a geodesic of M_2 . Notice that a geodesic mapping $\phi: M_1 \rightarrow M_2$ clearly also takes totally geodesic submanifolds of M_1 to totally geodesic submanifolds of M_2 .

As usual, we first consider S^n . We define the **central projection** ϕ of S^n to be the map which takes a point p in the open northern hemisphere S^{n+} of S^n to the intersection of $\mathbb{R}^n = \mathbb{R}^n \times \{1\} \subset \mathbb{R}^{n+1}$ with the straight line through p and the origin $0 \in \mathbb{R}^{n+1}$. It is clear that $\phi: S^{n+} \rightarrow \mathbb{R}^n$ is a geodesic mapping, since



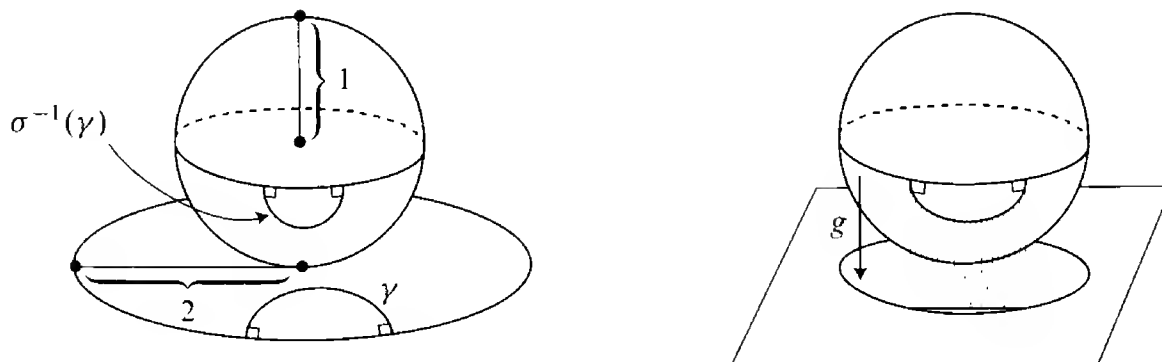
the geodesics of S^n are intersections of S^n with planes through the center of S^n . An exactly analogous construction works for $H^n \subset (\mathbb{R}^n, \langle \cdot, \cdot \rangle)$, except now we

obtain a map defined on all of H^n . We define $\phi: H^n \rightarrow \mathbb{R}^n$ to be the map which



takes $p \in H^n$ to the intersection of $\mathbb{R}^n = \{(1, a^1, \dots, a^n) \in \mathbb{R}^{n+1}\}$ with the straight line through p and $0 \in \mathbb{R}^{n+1}$. In this case, the image of H^n is the open ball in \mathbb{R}^n bounded by the intersection of \mathbb{R}^n with the cone $\sum_i (x^i)^2 = (x^0)^2$; thus $\phi(H^n)$ is the open ball $B^n(1)$ of radius 1.

We can also construct a geodesic mapping by using the model $(B^n, \langle \cdot, \cdot \rangle) = (B^n(2), \langle \cdot, \cdot \rangle)$. To do this, we regard S^n as the unit sphere tangent to $\mathbb{R}^n = \mathbb{R}^n \times \{0\}$ at 0. Then the stereographic projection σ from the north pole of S^n



takes the open southern hemisphere of S^n diffeomorphically onto $B^n(2)$. A

geodesic γ in $(B^n, \langle \cdot, \cdot \rangle)$ is a straight line or circle intersecting $S = \text{boundary } B^n$ orthogonally. It follows from the properties of stereographic projection that $\sigma^{-1}(\gamma)$ is a semi-circle intersecting the equator of S^n orthogonally. Now let $g: S^n \rightarrow \mathbb{R}^n$ be the orthogonal projection $g(x^1, \dots, x^{n+1}) = (x^1, \dots, x^n)$. Then g takes circles intersecting the equator of S^n orthogonally onto straight line segments of \mathbb{R}^n . So $g \circ \sigma^{-1}: B^n(2) \rightarrow B^n(1)$ is a geodesic mapping. Using these geodesic mappings $H^n \rightarrow B^n(1)$ and $B^n(2) \rightarrow B^n(1)$, it is not hard (Problems 9, 10) to describe an isometry between H^n and $B^n(2)$.

Naturally, the geodesic mapping $S^{n+} \rightarrow \mathbb{R}^n$ and $H^n \rightarrow B^n(1)$, together with the standard coordinate system on \mathbb{R}^n , lead to new coordinate systems for S^{n+} and H^n . In particular, the unit ball $B^n(1)$, together with the metric induced by the metric on H^n , is called the “projective model” of H^n .

We could calculate the form of the metric in these coordinate systems, and describe the geodesic spheres, horospheres, and equidistant hypersurfaces of H^n in the projective model, but we will never need to know any of this information. For us, the only important result will be the *existence* of the geodesic maps $S^{n+} \rightarrow \mathbb{R}^n$ and $H^n \rightarrow B^n(1)$. This is not surprising in view of the following:

2. THEOREM (BELTRAMI). If M is a connected Riemannian n -manifold such that every point has a neighborhood that can be mapped geodesically to \mathbb{R}^n , then M has constant curvature.

PROOF. The case $n \geq 3$ follows immediately from Theorem I-18; it is the case $n = 2$ which causes all the trouble. Note first that in the case of a surface, Lemma II.7-18 implies that the curvature K satisfies

$$\begin{aligned}
 R_{hijk} &= K(g_{hj}g_{ik} - g_{hk}g_{ij}) \\
 &\Downarrow \\
 (1) \quad R^l_{ijk} &= \sum_{h=1}^2 g^{lh} R_{hijk} = K(\delta_j^l g_{ik} - \delta_k^l g_{ij}).
 \end{aligned}$$

We will use the mapping in the hypothesis of the theorem to identify our neighborhood in M with an open set in \mathbb{R}^2 , on which we use the standard coordinate system (x^1, x^2) . Thus the metric $\sum_{i,j} g_{ij} dx^i \otimes dx^j$ has the same geodesics as the metric $\sum_{i,j} \delta_{ij} dx^i \otimes dx^j$. Since the Christoffel symbols for the latter metric are all zero, Proposition II.6-18 shows that the Christoffel symbols for g_{ij} satisfy

$$\Gamma_{jk}^i = \delta_j^i \omega_k + \delta_k^i \omega_j$$

for certain functions ω_i . Hence we have

$$(2) \quad \Gamma_{11}^2 = \Gamma_{22}^1 = 0, \quad \Gamma_{11}^1 = 2\Gamma_{12}^2, \quad \Gamma_{22}^2 = 2\Gamma_{21}^1.$$

From equation (1), and the formula (**) for R^l_{ijk} on pg. II.188, we find that

$$(3a) \quad Kg_{11} = (\Gamma_{12}^2)^2 - \frac{\partial}{\partial x^1} \Gamma_{12}^2 \quad (3b) \quad Kg_{12} = \Gamma_{21}^1 \Gamma_{12}^2 - \frac{\partial}{\partial x^2} \Gamma_{12}^2$$

$$(3c) \quad Kg_{22} = (\Gamma_{12}^1)^2 - \frac{\partial}{\partial x^2} \Gamma_{12}^1 \quad (3d) \quad Kg_{21} = \Gamma_{12}^2 \Gamma_{12}^1 - \frac{\partial}{\partial x^1} \Gamma_{12}^1.$$

Notice that equations (3b) and (3d) imply that

$$(4) \quad \frac{\partial}{\partial x^2} \Gamma_{12}^2 = \frac{\partial}{\partial x^1} \Gamma_{12}^1.$$

We also have

$$\begin{aligned} 2g_{12}\Gamma_{12}^2 &= g_{12}\Gamma_{11}^1 && \text{by (2)} \\ &= g_{12}\Gamma_{11}^1 + g_{22}\Gamma_{11}^2 && \text{by (2)} \\ &= [11, 2] \\ &= \frac{1}{2} \left(\frac{\partial g_{12}}{\partial x^1} + \frac{\partial g_{12}}{\partial x^1} - \frac{\partial g_{11}}{\partial x^2} \right) \\ &= \frac{\partial g_{12}}{\partial x^1} - \frac{1}{2} \frac{\partial g_{11}}{\partial x^2}. \end{aligned}$$

Subtracting this equation from

$$g_{11}\Gamma_{12}^1 + g_{12}\Gamma_{12}^2 = [12, 1] = \frac{1}{2} \frac{\partial g_{11}}{\partial x^2},$$

we obtain

$$g_{11}\Gamma_{12}^1 - g_{12}\Gamma_{12}^2 = \frac{\partial g_{11}}{\partial x^2} - \frac{\partial g_{12}}{\partial x^1}.$$

Multiplying by K , and using (3a) and (3b), we obtain

$$(5) \quad K \left(\frac{\partial g_{11}}{\partial x^2} - \frac{\partial g_{12}}{\partial x^1} \right) = \Gamma_{12}^2 \frac{\partial}{\partial x^2} \Gamma_{12}^2 - \Gamma_{12}^1 \frac{\partial}{\partial x^1} \Gamma_{12}^2.$$

Now differentiate (3a) with respect to x^2 , and subtract the result of differentiating (3b) with respect to x^1 . We obtain

$$\begin{aligned} g_{11} \frac{\partial K}{\partial x^2} - g_{12} \frac{\partial K}{\partial x^1} + K \left(\frac{\partial g_{11}}{\partial x^2} - \frac{\partial g_{12}}{\partial x^1} \right) \\ = 2\Gamma_{12}^2 \frac{\partial}{\partial x^2} \Gamma_{12}^2 - \Gamma_{21}^1 \frac{\partial}{\partial x^1} \Gamma_{12}^2 - \Gamma_{12}^2 \frac{\partial}{\partial x^1} \Gamma_{21}^1. \end{aligned}$$

Using (5) we have

$$g_{11} \frac{\partial K}{\partial x^2} - g_{12} \frac{\partial K}{\partial x^1} = \Gamma_{12}^2 \frac{\partial}{\partial x^2} \Gamma_{12}^2 - \Gamma_{12}^2 \frac{\partial}{\partial x^1} \Gamma_{21}^1.$$

Hence by (4) we have

$$g_{11} \frac{\partial K}{\partial x^2} - g_{12} \frac{\partial K}{\partial x^1} = 0.$$

Similarly,

$$-g_{12} \frac{\partial K}{\partial x^2} - g_{22} \frac{\partial K}{\partial x^1} = 0.$$

Since the determinant $g_{11}g_{22} - (g_{12})^2 \neq 0$, this implies that $\partial K/\partial x^1 = 0$ and $\partial K/\partial x^2 = 0$. ♦

B. CURVES IN A RIEMANNIAN MANIFOLD

Before investigating general submanifolds of a Riemannian manifold, we will consider the special case of 1-dimensional submanifolds, which works out quite differently than all other cases. Our aim is not to obtain any particularly startling theorems about curves in Riemannian manifolds, but merely to show briefly how the Serret-Frenet formulas of Chapter II.1 generalize; along the way we will derive a few results which are needed to discuss higher dimensions.

Consider a Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$, and an arclength parameterized curve $c: [a, b] \rightarrow N$. We use N for the ambient manifold to conform with the notation to be used in the general case of a submanifold $M \subset N$. For consistency of notation, we also use ∇' for the covariant derivative in N , even though there will be no occasion to consider the covariant derivative ∇ in the 1-dimensional manifold $c([a, b])$. We will let $\mathbf{v}_1 = c'$ denote the unit tangent vector of c . Since $\langle \mathbf{v}_1, \mathbf{v}_1 \rangle = 1$ we have

$$0 = \frac{d}{ds} \langle \mathbf{v}_1(s), \mathbf{v}_1(s) \rangle = 2 \left\langle \mathbf{v}_1(s), \frac{D' \mathbf{v}_1(s)}{ds} \right\rangle.$$

We define the first “curvature function” κ_1 of c by

$$\kappa_1(s) = \left| \frac{D' \mathbf{v}_1(s)}{ds} \right|,$$

and if $\kappa_1(s) \neq 0$ for all s we set

$$\mathbf{v}_2(s) = \kappa_1(s)^{-1} \cdot \frac{D' \mathbf{v}_1(s)}{ds},$$

so that \mathbf{v}_2 is a unit vector field along c which is everywhere perpendicular to \mathbf{v}_1 . We then have the “Frenet formula”

$$(F_1) \quad \frac{D'\mathbf{v}_1(s)}{ds} = \kappa_1(s)\mathbf{v}_2(s).$$

Now

$$\langle \mathbf{v}_2, \mathbf{v}_2 \rangle = 1 \implies \left\langle \mathbf{v}_2(s), \frac{D'\mathbf{v}_2(s)}{ds} \right\rangle = 0.$$

Moreover,

$$\begin{aligned} \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0 &\implies 0 = \left\langle \frac{D'\mathbf{v}_1(s)}{ds}, \mathbf{v}_2(s) \right\rangle + \left\langle \mathbf{v}_1(s), \frac{D'\mathbf{v}_2(s)}{ds} \right\rangle \\ &= \kappa_1(s) + \left\langle \mathbf{v}_1(s), \frac{D'\mathbf{v}_2(s)}{ds} \right\rangle \quad \text{by (F}_1\text{)}. \end{aligned}$$

This implies that

$$\frac{D'\mathbf{v}_2(s)}{ds} = -\kappa_1(s)\mathbf{v}_1(s) + \text{vector perpendicular to } \mathbf{v}_1(s) \text{ and } \mathbf{v}_2(s).$$

We define the second “curvature function” κ_2 by

$$\kappa_2(s) = \left| \frac{D'\mathbf{v}_2(s)}{ds} + \kappa_1(s)\mathbf{v}_1(s) \right|,$$

and if $\kappa_2(s) \neq 0$ for all s , we set

$$\mathbf{v}_3(s) = \kappa_2(s)^{-1} \left[\frac{D'\mathbf{v}_2(s)}{ds} + \kappa_1(s)\mathbf{v}_1(s) \right],$$

so that \mathbf{v}_3 is a unit vector field along c which is everywhere perpendicular to \mathbf{v}_1 and \mathbf{v}_2 . We then have

$$(F_2) \quad \frac{D'\mathbf{v}_2(s)}{ds} = -\kappa_1(s)\mathbf{v}_1(s) + \kappa_2(s)\mathbf{v}_3(s).$$

Now suppose, inductively, that for $j \leq m = \dim N$ we have orthonormal vector fields $\mathbf{v}_1, \dots, \mathbf{v}_j$ along c and nowhere zero curvature functions $\kappa_1, \dots, \kappa_{j-1}$ such that

$$(F_1) \quad \frac{D'\mathbf{v}_1(s)}{ds} = \kappa_1(s)\mathbf{v}_2(s)$$

$$(F_2) \quad \frac{D'\mathbf{v}_2(s)}{ds} = -\kappa_1(s)\mathbf{v}_1(s) + \kappa_2(s)\mathbf{v}_3(s)$$

\vdots

$$(F_{j-1}) \quad \frac{D'\mathbf{v}_{j-1}(s)}{ds} = -\kappa_{j-2}(s)\mathbf{v}_{j-2}(s) + \kappa_{j-1}(s)\mathbf{v}_j(s).$$

Then

$$\langle \mathbf{v}_j, \mathbf{v}_j \rangle = 1 \implies \left\langle \frac{D'\mathbf{v}_j(s)}{ds}, \mathbf{v}_j(s) \right\rangle = 0,$$

while for $i < j$ we have

$$\begin{aligned} \langle \mathbf{v}_j, \mathbf{v}_i \rangle = 0 &\implies \left\langle \mathbf{v}_i(s), \frac{D'\mathbf{v}_j(s)}{ds} \right\rangle = - \left\langle \frac{D'\mathbf{v}_i(s)}{ds}, \mathbf{v}_j \right\rangle \\ &= \begin{cases} 0 & i \neq j-1 \\ -\kappa_{j-1}(s) & i = j-1. \end{cases} \end{aligned}$$

Hence

$$\begin{aligned} (*) \quad \frac{D'\mathbf{v}_j(s)}{ds} &= -\kappa_{j-1}(s)\mathbf{v}_{j-1}(s) \\ &\quad + \text{vector perpendicular to } \mathbf{v}_1(s), \dots, \mathbf{v}_j(s). \end{aligned}$$

If $j < m$ we set

$$\kappa_j(s) = \left| \frac{D'\mathbf{v}_j(s)}{ds} + \kappa_{j-1}(s)\mathbf{v}_{j-1}(s) \right|,$$

and if $\kappa_j(s) \neq 0$ for all s we set

$$\mathbf{v}_{j+1}(s) = \kappa_j(s)^{-1} \cdot \left[\frac{D'\mathbf{v}_j(s)}{ds} + \kappa_{j-1}(s)\mathbf{v}_{j-1}(s) \right].$$

We then have

$$(F_j) \quad \frac{D'\mathbf{v}_j(s)}{ds} = -\kappa_{j-1}(s)\mathbf{v}_{j-1}(s) + \kappa_j(s)\mathbf{v}_{j+1}(s).$$

If $j = m$, then only the zero vector is perpendicular to $\mathbf{v}_1(s), \dots, \mathbf{v}_m(s)$, so equation (*) becomes

$$(F_m) \quad \frac{D'\mathbf{v}_m(s)}{ds} = -\kappa_{m-1}(s)\mathbf{v}_{m-1}(s).$$

It is easy to see that we have equations (F₁) to (F_{j-1}) with nowhere zero functions $\kappa_1, \dots, \kappa_{j-1}$ if and only if

$$c'(s), \quad \frac{D'c'(s)}{ds}, \quad \dots, \quad \frac{D'^{(j-1)}c'(s)}{ds^{j-1}}$$

are everywhere linearly independent; the vector fields $\mathbf{v}_1, \dots, \mathbf{v}_j$ along c are then precisely the result of applying the Gram-Schmidt orthonormalization

process to these vectors. If $D'^{(j)}c'(s)/ds^j$ is everywhere linearly dependent on $c'(s), \dots, D'^{(j-1)}c'(s)/ds^{j-1}$, then the function κ_j will be everywhere 0, and we cannot define \mathbf{v}_{j+1} , but we can write instead

$$(F'_j) \quad \frac{D'\mathbf{v}_j(s)}{ds} = -\kappa_{j-1}(s)\mathbf{v}_{j-1}(s).$$

[Note, in particular, that (F'_m) is just (F_m) .] As in the theory of curves in \mathbb{R}^3 , we consider only intervals where a set of equations $(F_1), \dots, (F_{j-1}), (F'_j)$ holds for some $j \leq m$. In other words, we assume that $\kappa_1, \dots, \kappa_{j-1}$ are nowhere zero, while κ_j is identically zero. We call $\mathbf{v}_1, \dots, \mathbf{v}_j$ the “Frenet frame” for c . The subspace of $N_{c(s)}$ spanned by $\mathbf{v}_1(s)$ and $\mathbf{v}_i(s)$ is sometimes called the $(i-1)^{\text{st}}$ **osculating plane** of c at s .

Notice that once we have $c'(s), \dots, D'^{(m-2)}c'(s)/ds^{m-2}$ linearly independent, so that $\mathbf{v}_1, \dots, \mathbf{v}_{m-1}$ are defined, then there are only two possible choices for each $\mathbf{v}_m(s)$. Having made a choice of $\mathbf{v}_m(a)$, there is then a unique continuous way of choosing $\mathbf{v}_m(s)$ for all $s \in [a, b]$. We still have equations (F_1) to (F_m) , but now the function κ_{m-1} might take on negative values, whereas all other κ_j , being non-zero norms, are everywhere positive. The particularly interesting situation occurs when N is oriented. Then we define $\mathbf{v}_m(s)$ to be the unique unit vector in $N_{c(s)}$ orthogonal to $\mathbf{v}_1(s), \dots, \mathbf{v}_{m-1}(s)$ such that $(\mathbf{v}_1(s), \dots, \mathbf{v}_m(s))$ is positively oriented [equivalently, we can define $\mathbf{v}_m(s) = \mathbf{v}_1(s) \times \dots \times \mathbf{v}_{m-1}(s)$, where the cross-product is determined by the metric and the orientation (see Problem 11)]. For curves in \mathbb{R}^3 this is precisely how we defined the binormal $\mathbf{b} = \mathbf{v}_3$, and obtained the torsion $\tau = \kappa_2$ which could be positive, negative, or zero. When we apply this procedure to arclength parameterized curves c in an oriented 2-dimensional Riemannian manifold, we obtain an everywhere defined curvature κ_1 , whose values may be positive, negative, or zero. Clearly, κ_1 is just the geodesic curvature κ_g defined previously.

In the next theorem we will, for simplicity, ignore these refinements and consider only curves with $c'(s), \dots, D'^{(m-1)}c'(s)/ds^{m-1}$ everywhere linearly independent. Readers may sort out for themselves the details which have to be changed when N is oriented and we allow κ_{m-1} to take on non-positive values.

3. THEOREM. (I) Let $c, \bar{c}: [a, b] \rightarrow N^m$ be arclength parameterized curves with nowhere zero curvature functions $\kappa_1, \dots, \kappa_{m-1}$ and $\bar{\kappa}_1, \dots, \bar{\kappa}_{m-1}$, respectively; and Frenet frames $\mathbf{v}_1, \dots, \mathbf{v}_m$ and $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_m$, respectively. Suppose that $\kappa_i = \bar{\kappa}_i$ for $1 \leq i \leq m-1$, and that

$$c(a) = \bar{c}(a) \quad \text{and} \quad \mathbf{v}_i(a) = \bar{\mathbf{v}}_i(a) \quad \text{for } i = 1, \dots, m.$$

Then $c = \bar{c}$.

(2) Let N^m be complete, let $\kappa_1, \dots, \kappa_{m-1}: [a, b] \rightarrow \mathbb{R}$ be everywhere positive continuous functions, and let $\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_m$ be an orthonormal basis for some N_p . Then there is an arclength parameterized curve $c: [a, b] \rightarrow N$ with $c(a) = p$ whose curvature functions are $\kappa_1, \dots, \kappa_{m-1}$ and whose Frenet frame $\mathbf{v}_1, \dots, \mathbf{v}_m$ satisfies $\mathbf{v}_i(a) = \hat{\mathbf{v}}_i$ for $1 \leq i \leq m$.

PROOF. To prove (1) it clearly suffices (by a least upper bound argument) to show that $c(s) = \bar{c}(s)$ for s sufficiently close to a . So we might as well assume that $M = \mathbb{R}^m$, with some metric $\sum_{i,j} g_{ij} dx^i \otimes dx^j$, where x^1, \dots, x^m is the standard coordinate system for \mathbb{R}^m . Let $v_i(s) \in \mathbb{R}^m$ be the vector representing $\mathbf{v}_i(s)$ when we identify tangent vectors of \mathbb{R}^m with elements of \mathbb{R}^m in the usual way. We thus have $m+1$ functions

$$c, v_1, \dots, v_m: [a, b] \rightarrow \mathbb{R}^m.$$

We will also let $Dv_i(s)/ds$ be the vector representing $D\mathbf{v}_i(s)/ds$ when we identify tangent vectors of \mathbb{R}^m with elements of \mathbb{R}^m . The formula on pg. II.232 shows that $Dv_j(s)/ds$ can be written in terms of

$$c(s), c'(s), v_j(s), v_j'(s), \quad \text{i.e.,} \quad c(s), v_1(s), v_j(s), v_j'(s).$$

Each Frenet equation (F_j) then gives us an equation

$$(E_j) \quad v_j'(s) = F_j(c(s), v_1(s), \dots, v_m(s)).$$

We also have the equation

$$(E_0) \quad c'(s) = v_1(s).$$

So the equations $(E_0), (E_1), \dots, (E_m)$ gives us an equation

$$(*) \quad \alpha'(s) = F(\alpha(s))$$

for the function $\alpha = (c, v_1, \dots, v_m)$. The function F depends only on $\kappa_1, \dots, \kappa_{m-1}$ (and the Christoffel symbols).

For the function $\bar{\alpha} = (\bar{c}, \bar{v}_1, \dots, \bar{v}_m)$ there is a similar equation

$$\bar{\alpha}'(s) = \bar{F}(\bar{\alpha}(s)).$$

Moreover, since $\bar{\kappa}_i = \kappa_i$ for all i , the function \bar{F} is exactly the same as F . Now by hypothesis, the functions α and $\bar{\alpha}$ are equal at a , so by uniqueness of solutions of $(*)$ we have $\alpha = \bar{\alpha}$. Hence $c = \bar{c}$.

To prove (2) we first show that the desired curve c can be defined on some interval $[a, a + \varepsilon]$. Again we may clearly assume that $M = \mathbb{R}^m$. Now we can solve equation (*) on some interval $[a, a + \varepsilon]$, with any given initial conditions. This gives us a curve $c: [a, a + \varepsilon] \rightarrow M$ and vector fields $c' = \mathbf{v}_1, \dots, \mathbf{v}_m$ along c satisfying the Frenet formulas (F_1) to (F_m) , with

$$c(a) = p \quad \text{and} \quad \mathbf{v}_i(a) = \mathring{\mathbf{v}}_i, \quad 1 \leq i \leq m.$$

We have

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle'(s) = \left\langle \frac{D\mathbf{v}_i(s)}{ds}, \mathbf{v}_j(s) \right\rangle + \left\langle \mathbf{v}_i(s), \frac{D\mathbf{v}_j(s)}{ds} \right\rangle;$$

using the formulas (F_1) to (F_m) , we find that this is zero. Since $\{\mathbf{v}_i(a)\} = \{\mathring{\mathbf{v}}_i\}$ is orthonormal, $\{\mathbf{v}_i\}$ must therefore be orthonormal everywhere. So $\mathbf{v}_1, \dots, \mathbf{v}_m$ is the Frenet frame of c , and the κ_i are its curvatures.

In order to extend c to all of $[a, b]$, we first consider the equation (*) once again. If we choose our initial point p to lie in some compact set, then there will be $\varepsilon > 0$ with the property that for any orthonormal $\{\mathring{\mathbf{v}}_i\}$ at any such p , there is a curve $c: [a, a + \varepsilon] \rightarrow N$ with curvature functions κ_i on $[a, a + \varepsilon]$, whose Frenet frame $\mathbf{v}_1, \dots, \mathbf{v}_m$ satisfies $\mathbf{v}_i(a) = \mathring{\mathbf{v}}_i$. The size of ε will depend on bounds for F , and hence only on bounds for the κ_i , as well as bounds for the Christoffel symbols Γ_{ij}^k . It is thus clear that for every point $q \in M$ there is $\delta(q) > 0$ with the following property:

- (**) If $d(p, q) < \delta(q)$ and $\{\mathring{\mathbf{v}}_i\}$ is an orthonormal basis for M_p , then for any a' with $a < a' < b$ there is a curve $c: [a', \min(a' + \delta(q), b)] \rightarrow M$ with curvature functions κ_i on this interval, whose Frenet frame $\mathbf{v}_1, \dots, \mathbf{v}_m$ satisfies $\mathbf{v}_i(a') = \mathring{\mathbf{v}}_i$.

Now by a least upper bound argument it clearly suffices to show that the curve $c: [a, a + \varepsilon] \rightarrow M$, with Frenet frame $\mathbf{v}_1, \dots, \mathbf{v}_m$, can always be extended to the closed interval $[a, a + \varepsilon]$. The curve c is parameterized by arclength (since $\mathbf{v}_1 = c'$ has length 1), so for all $a < a' < a + \varepsilon$ we have $d(c(a), c(a')) \leq \text{length of } c \text{ on } [a, a'] = a' - a < \varepsilon$. So the image of c on $[a, a + \varepsilon]$ lies in some compact subset K of the complete manifold M . Thus there is $\delta > 0$ which will serve as $\delta(q)$ in (**) for all $q \in K$. Now choose $a < a' < a + \varepsilon$, so that $(a + \varepsilon) - a' < \delta$, and find the curve \bar{c} with curvature functions κ_i , whose Frenet frame $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_m$ satisfies $\bar{\mathbf{v}}_i(a') = \mathbf{v}_i(a')$. The curve \bar{c} is defined at least on $[a', \varepsilon]$ by (**), and c followed by \bar{c} is an extension of c at least as far as $c + \varepsilon$. ♦

4. COROLLARY. Let N be a complete m -dimensional Riemannian manifold of constant curvature K_0 . Let $c, \bar{c}: [a, b] \rightarrow N$ be arclength parameterized curves with nowhere zero curvature functions $\kappa_1, \dots, \kappa_{m-1}$ and $\bar{\kappa}_1, \dots, \bar{\kappa}_{m-1}$, respectively. If $\kappa_i = \bar{\kappa}_i$ for $1 \leq i \leq m-1$, then there is a unique isometry $A: N \rightarrow N$ such that $\bar{c} = A \circ c$.

PROOF. Left to the reader. ♦

We also want to consider curves with $\kappa_1, \dots, \kappa_{j-1}$ nowhere zero, but κ_j identically zero, for some $j \leq m-1$. In Chapter II.1 we found that curves with $\kappa = 0$ are straight lines, while curves with $\tau = 0$ lie in a plane. The generalization for curves in \mathbb{R}^m is the following.

5. THEOREM. Let $c: [a, b] \rightarrow \mathbb{R}^m$ be an arclength parameterized curve with $\kappa_1, \dots, \kappa_{j-1}$ nowhere zero, and κ_j everywhere zero. Then c lies in some j -dimensional plane in \mathbb{R}^m .

PROOF. Let $\mathbf{v}_1, \dots, \mathbf{v}_j$ be the Frenet frame for c , and let $\Delta(s) \subset \mathbb{R}^m_{c(s)}$ be the j -dimensional subspace of $\mathbb{R}^m_{c(s)}$ spanned by $\mathbf{v}_1(s), \dots, \mathbf{v}_j(s)$. We claim that all $\Delta(s)$ are parallel (considered as j -dimensional planes in \mathbb{R}^m). To prove this, we note that since $D'\mathbf{v}_i(s)/ds$ is just $\mathbf{v}_i'(s)$ in \mathbb{R}^m , the Frenet equations $(F_1), \dots, (F_{j-1}), (F'_j)$ show that each $\mathbf{v}_i'(s)$ is a linear combination of certain of the $\mathbf{v}_i(s)$,

$$\mathbf{v}_i'(s) = \sum_{t=1}^j a_{it}(s) \mathbf{v}_t(s).$$

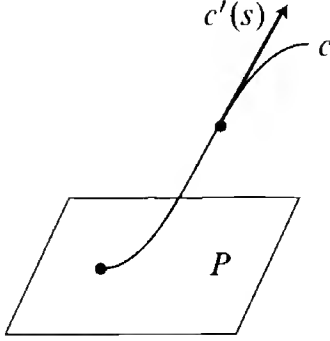
So if \mathbf{w} is a parallel vector field on \mathbb{R}^m (that is, if for some $w \in \mathbb{R}^m$ we have $\mathbf{w}(p) = w_p$ for all p), then

$$(*) \quad \frac{d}{ds} \langle \mathbf{v}_i(s), \mathbf{w}(c(s)) \rangle = \langle \mathbf{v}_i'(s), \mathbf{w}(c(s)) \rangle = \sum_{t=1}^j a_{it}(s) \langle \mathbf{v}_t(s), \mathbf{w}(c(s)) \rangle.$$

By uniqueness of solutions of the system $(*)$, we see that if all $\langle \mathbf{v}_i(a), \mathbf{w}(c(a)) \rangle = 0$, then all $\langle \mathbf{v}_i(s), \mathbf{w}(c(s)) \rangle = 0$ for all s . In other words, $\Delta(s)$ is always orthogonal to the same vectors as $\Delta(a)$. Hence $\Delta(s)$ is parallel to $\Delta(a)$. Our result now follows from

6. LEMMA. Let $c: [a, b] \rightarrow \mathbb{R}^m$ be an immersed curve, and for each s let $\Delta(s) \subset \mathbb{R}^m_{c(s)}$ be a j -dimensional subspace of $\mathbb{R}^m_{c(s)}$ with $c'(s) \in \Delta(s)$. Suppose that all $\Delta(s)$ are parallel. Then c is a curve in some j -dimensional plane $P \subset \mathbb{R}^m$, and P is just $\exp(\Delta(s))$ for any s .

PROOF. Let $P = \Delta(a)$, considered as a j -dimensional plane in \mathbb{R}^m . Without loss of generality we may assume that P is parallel to the (x^1, \dots, x^j) -plane. If c does not lie entirely in P , then by the mean value theorem some tangent vector $c'(s)$ has a non-zero k^{th} component for some $k > j$. But this is impossible,



since $c'(s) \in \Delta(s)$ and $\Delta(s)$ is parallel to $P = \Delta(a)$. So c lies in P . Since each $\Delta(s)$ is parallel to $P = \Delta(a)$ and also contains the point $c(s) \in P$, each $\Delta(s)$ must equal P when $\Delta(s)$ is considered as a j -dimensional plane in \mathbb{R}^m . In other words, $P = \exp(\Delta(s))$. ♦

As soon as we try to replace \mathbb{R}^m in Theorem 5 with a manifold N of constant curvature, we find that the proof of Lemma 6 doesn't generalize at all. However, the result is still true, and we will give two proofs, exploiting two different descriptions of constant curvature manifolds. First consider a curve $c: [a, b] \rightarrow N$ in any Riemannian manifold N , and suppose that for each s we have a j -dimensional subspace $\Delta(s) \subset N_{c(s)}$, so that Δ is a “distribution along c ”. Let $\tau_s: N_{c(a)} \rightarrow N_{c(s)}$ be the parallel translation along c from $c(a)$ to $c(s)$. We say that Δ is **parallel along c** if $\tau_s(\Delta(a)) = \Delta(s)$ for all s . Suppose that Δ is parallel along c and that V is a smooth vector field along c belonging to Δ (that is, $V(s) \in \Delta(s)$ for all s). Proposition II.6-3 immediately shows that

$$\frac{D'V(s)}{ds} \in \Delta(s) \quad \text{for all } s,$$

so that $D'V(s)/ds$ also belongs to Δ . We need the converse assertion also.

7. PRE-LEMMA. Let $c: [a, b] \rightarrow N$ be a curve in a Riemannian manifold N , and let Δ be a smooth j -dimensional distribution along c . Suppose that $D'V(s)/ds$ belongs to Δ whenever V is a smooth vector field belonging to Δ . Then Δ is parallel along c .

PROOF. Choose everywhere linearly independent smooth vector fields V_1, \dots, V_j along c belonging to Δ . By hypothesis, we can write

$$(1) \quad \frac{D' V_i(s)}{ds} = \sum_{t=1}^j a_{it}(s) V_t(s)$$

for certain smooth functions a_{it} . We claim that there are functions $b_{\lambda i}$, with arbitrary initial conditions $b_{\lambda i}(a)$, such that

$$(2) \quad \frac{D'}{ds} \sum_{\lambda=1}^j b_{\lambda i}(s) V_\lambda(s) = 0, \quad i = 1, \dots, j.$$

In fact, equation (2) is equivalent to

$$\begin{aligned} 0 &= \sum_{\lambda=1}^j b_{\lambda i}'(s) V_\lambda(s) + \sum_{\lambda=1}^j b_{\lambda i}(s) \frac{D' V_\lambda(s)}{ds} \\ &= \sum_{t=1}^j b_{it}'(s) V_t(s) + \sum_{\lambda, t=1}^j b_{\lambda i}(s) a_{t\lambda}(s) V_t(s) \quad \text{by (1),} \end{aligned}$$

and hence to

$$(3) \quad b_{it}'(s) = \sum_{\lambda=1}^j a_{t\lambda}(s) b_{\lambda i}(s), \quad i = 1, \dots, j.$$

Since (3) is a linear equation, we can solve it on the whole interval $[a, b]$, with arbitrary initial conditions. Choose the initial conditions $b_{\lambda i}(a) = \delta_{\lambda i}$, and set

$$W_i(s) = \sum_{\lambda=1}^j b_{\lambda i}(s) V_\lambda(s).$$

Then the vector fields W_i along c are parallel, by (2), and linearly independent at a , hence linearly independent everywhere. So the $W_i(s)$ span $\Delta(s)$ for all s , which shows that Δ is parallel along c . ♦

8. LEMMA. Let N be a manifold of constant curvature K_0 . Let $c: [a, b] \rightarrow N$ be an immersed curve, and let Δ be a j -dimensional distribution along c such that $c'(s) \in \Delta(s)$ for all s . Suppose that Δ is parallel along c . Then c lies in some j -dimensional totally geodesic submanifold $P \subset N$, and $\exp(\Delta(s)) \subset P$ for all s .

FIRST PROOF. It is easy to see that this result is essentially a local one, so without loss of generality we take N to be the complete simply-connected manifold of constant curvature K_0 . Consider first the case $K_0 > 0$, so that $N = S^m(K_0) \subset \mathbb{R}^{m+1}$. We can then consider $V(s) \subset \mathbb{R}^{m+1}_{c(s)}$. We denote the covariant derivatives in N and \mathbb{R}^{m+1} by ∇' and ∇ , respectively, and we will let \mathbf{v} be a unit normal field on $N = S^m(K_0)$.

Let V be a vector field along c which belongs to Δ , so that $D'V/ds$ also belongs to Δ , since Δ is parallel along c . If \mathbf{D}'/ds denotes the covariant derivative along c in \mathbb{R}^{n+1} , then Corollary 1-2 gives

$$\mathbf{T} \left(\frac{\mathbf{D}'V(s)}{ds} \right) = \frac{D'V(s)}{ds} \in \Delta(s).$$

Thus we see that

$$(1) \quad V(s) \in \Delta(s) \text{ for all } s \implies \frac{\mathbf{D}'V(s)}{ds} \in \Delta(s) + \mathbb{R} \cdot \mathbf{v}(c(s)) \text{ for all } s.$$

On the other hand, we also have

$$(2) \quad \begin{aligned} \frac{\mathbf{D}'\mathbf{v}(c(s))}{ds} &= \nabla'_{c'(s)}\mathbf{v} = (\text{constant}) \cdot c'(s), \\ &\text{since all points of } N \text{ are umbilics} \\ &\in \Delta(s), \quad \text{by assumption on } \Delta. \end{aligned}$$

Now let

$$\mathbf{\Delta}(s) = \Delta(s) \oplus \mathbb{R} \cdot \mathbf{v}(c(s)) \in \mathbb{R}^{m+1}_{c(s)}.$$

From (1) and (2) we see that for a vector field W along c in \mathbb{R}^{m+1} we have

$$W(s) \in \mathbf{\Delta}(s) \text{ for all } s \implies \frac{\mathbf{D}'W(s)}{ds} \in \mathbf{\Delta}(s) \text{ for all } s.$$

By our prelemmanary remark we see that $\mathbf{\Delta}$ is parallel along c in \mathbb{R}^{m+1} . So Lemma 6 shows that c lies in some $(j+1)$ -dimensional plane $P \subset \mathbb{R}^{m+1}$, and $P = \exp(\mathbf{\Delta}(s))$ for all s . Since $\mathbf{v}(c(s)) \in \mathbf{\Delta}(s)$, the plane P must pass through the origin $0 \in \mathbb{R}^{n+1}$. Hence c is contained in $P \cap S^m(K_0)$, which is a j -dimensional totally geodesic subspace of $S^m(K_0)$. Clearly, we also have $\exp(\Delta(s)) \subset P \cap S^m(K_0)$ for all s .

The case $K_0 < 0$ can be proved similarly, taking N to be $H^m(K_0)$, considered as a subset of \mathbb{R}^{m+1} with the Lorentzian metric.

SECOND PROOF. As the result is essentially local, we can assume that we have a geodesic mapping $\phi: N \rightarrow \mathbb{R}^m$. If $\bar{\nabla}$ denotes covariant differentiation on \mathbb{R}^m , then Proposition II.6-18 shows that there is a 1-form ω on \mathbb{R}^m with

$$(1) \quad \bar{\nabla}_{\phi_* X} \phi_* Y - \phi_*(\nabla'_X Y) = \omega(\phi_* X) \cdot \phi_* Y + \omega(\phi_* Y) \cdot \phi_* X.$$

Let

$$\gamma(s) = \phi(c(s))$$

$$\bar{\Delta}(s) = \phi_* \Delta(s) \subset \mathbb{R}^m_{\gamma(s)}.$$

Then we have

$$(2) \quad \gamma'(s) = \phi_* c'(s) \in \phi_* \Delta(s) = \bar{\Delta}(s).$$

If V is a vector field along c with $V(s) \in \Delta(s)$ for all s , and hence $D'V(s)/ds \in \Delta(s)$ for all s , the equation (1) implies that

$$\begin{aligned} \frac{\bar{D}\phi_* V(s)}{ds} &= \phi_* \left(\frac{D'V(s)}{ds} \right) + \omega(\gamma'(s)) \cdot \phi_* V(s) + \omega(\phi_* V(s)) \cdot \gamma'(s) \\ &\in \bar{\Delta}(s), \quad \text{by (2).} \end{aligned}$$

In other words, if W is a vector field along γ with $W(s) \in \bar{\Delta}(s)$ for all s , then also $\bar{D}W(s)/ds \in \bar{\Delta}(s)$ for all s . Once again, this implies that $\bar{\Delta}$ is parallel along γ . So Lemma 6 implies that γ lies in some j -dimensional plane $P \subset \mathbb{R}^m$. Then c lies in $\phi^{-1}(P)$, which is totally geodesic j -dimensional submanifold of N . Clearly, we also have $\exp(\Delta(s)) \subset \phi^{-1}(P)$ for all s . ♦

9. THEOREM. Let N be a manifold of constant curvature K_0 . Let $c: [a, b] \rightarrow N$ be an arclength parameterized curve with $\kappa_1, \dots, \kappa_{j-1}$ nowhere zero, and κ_j everywhere zero. Then c lies in some j -dimensional totally geodesic submanifold of N .

PROOF. Let $\mathbf{v}_1, \dots, \mathbf{v}_j$ be the Frenet frame for c , and let $\Delta(s) \subset M_{c(s)}$ be the subspace spanned by $\mathbf{v}_1(s), \dots, \mathbf{v}_j(s)$. The argument in the proof of Theorem 5 shows generally that Δ is parallel along c . So our result follows from Lemma 8. ♦

10. COROLLARY. Let N^m be a complete manifold of constant curvature K_0 . Let $c, \bar{c}: [a, b] \rightarrow N$ be arclength parameterized curves with $\kappa_1, \dots, \kappa_{j-1}$ and $\bar{\kappa}_1, \dots, \bar{\kappa}_{j-1}$ nowhere zero, and κ_j and $\bar{\kappa}_j$ everywhere zero. If $\kappa_i = \bar{\kappa}_i$ for $1 \leq i \leq j-1$, then there is an isometry $A: N \rightarrow N$ such that $\bar{c} = A \circ c$. The group of all such isometries is isomorphic to the orthogonal group $O(m-j-1)$.

PROOF. Left to the reader. ♦

For later use, we note a consequence of Lemma 8 for higher dimensional submanifolds M of N .

11. COROLLARY. Let N be a manifold of constant curvature K_0 . Let M be a connected submanifold immersed in N , and let Δ be a j -dimensional distribution along M such that $M_p \subset \Delta(p)$ for all $p \in M$. Suppose that Δ is parallel along every curve c in M . Then M lies in some j -dimensional totally geodesic submanifold $P \subset N$, and $\exp(\Delta(p)) \subset P$ for all $p \in M$.

PROOF. Choose a point $p_0 \in M$, and let P be the largest j -dimensional totally geodesic submanifold of N with $P \supset \exp(\Delta(p_0))$. For any $p \in M$, choose a curve $c: [0, 1] \rightarrow M$ with $c(0) = p_0$ and $c(1) = p$. Lemma 8, applied to the distribution $s \mapsto \Delta(c(s))$ along c , implies that c lies in some j -dimensional totally geodesic submanifold $P' \subset M$, and $\exp(\Delta(c(s))) \subset P'$ for all s . Applying this for $s = 0$, we see that $P' \subset P$. Hence $p \in P$, and also $\exp(\Delta(p)) = \exp(\Delta(c(1))) \subset P' \subset P$. ♦

C. THE FUNDAMENTAL EQUATIONS FOR SUBMANIFOLDS

In Chapter I, we considered a submanifold M^n of a Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$ with $i: M \rightarrow N$ the inclusion map. For each $p \in M$, we have $N_p = M_p \oplus M_p^\perp$, and we used this decomposition to define two projections, $\mathbb{T}: N_p \rightarrow M_p$ and $\mathbb{L}: N_p \rightarrow M_p^\perp$. For vector fields X and Y tangent along M we wrote

$$\nabla'_{X_p} Y = \mathbb{T}(\nabla'_{X_p} Y) + \mathbb{L}(\nabla'_{X_p} Y)$$

where ∇' is the covariant differentiation in N , and we showed that $\mathbb{T}(\nabla'_{X_p} Y) = \nabla_{X_p} Y$, where ∇ is the covariant differentiation in M determined by the metric $i^*\langle \cdot, \cdot \rangle$, while $\mathbb{L}(\nabla'_{X_p} Y) = s(X_p, Y_p)$ is symmetric in X_p and Y_p (and independent of the extension Y of Y_p). This gave us

The Gauss Formulas: $\nabla'_{X_p} Y = \nabla_{X_p} Y + s(X_p, Y_p)$

and we then derived

Gauss' Equation:

$$\begin{aligned} \langle R'(X, Y)Z, W \rangle &= \langle R(X, Y)Z, W \rangle \\ &\quad + \langle s(X, Z), s(Y, W) \rangle - \langle s(Y, Z), s(X, W) \rangle \end{aligned}$$

for all tangent vectors $X, Y, Z, W \in M_p$. (For convenience we will often use X, Y, Z, \dots without subscripts to denote vectors as well as vector fields.)

For a *hypersurface* with a unit normal field ν , we showed that $\nabla'_{X_p} \nu \in M_p$, and we determined this vector explicitly by

$$\text{The Weingarten Equations:} \quad \langle \nabla'_{X_p} \nu, Y_p \rangle = -\langle \nu, s(X_p, Y_p) \rangle.$$

Defining a tensor Π by $s(X_p, Y_p) = \Pi(X_p, Y_p) \cdot \nu(p)$, we then derived the

Codazzi-Mainardi Equations:

$$\langle R'(X, Y)Z, \nu \rangle = (\nabla_X \Pi)(Y, Z) - (\nabla_Y \Pi)(X, Z).$$

Now we want to consider a submanifold of arbitrary codimension. We define the **normal bundle** $\text{Nor } M$ of M in N to be

$$\text{Nor } M = \bigcup_{p \in M} M_p^\perp,$$

and we define the projection map

$$\varpi: \text{Nor } M \rightarrow M$$

to be the one which takes all vectors in M_p^\perp to p . Thus (compare pg. I.344) $\varpi: \text{Nor } M \rightarrow M$ is a vector bundle whose fibre $\varpi^{-1}(p)$ over p is M_p^\perp . A section ξ of E is a map with $\xi(p) \in M_p^\perp$ for all p , in other words, a normal vector field along M .

Unlike the case of a hypersurface, it is no longer true that $\nabla'_{X_p} \xi \in M_p$, even if ξ always has length 1, so we will look at the general decomposition

$$\nabla'_{X_p} \xi = \mathsf{T}(\nabla'_{X_p} \xi) + \mathsf{L}(\nabla'_{X_p} \xi).$$

The tangential component is just as nice as in the case of hypersurfaces:

12. PROPOSITION. If ξ is a section of the normal bundle of M , and $X_p \in M_p$, then the vector $\mathsf{T}(\nabla'_{X_p} \xi) \in M_p$ satisfies

$$\langle \mathsf{T}(\nabla'_{X_p} \xi), Y_p \rangle = \langle \nabla'_{X_p} \xi, Y_p \rangle = -\langle \xi(p), s(X_p, Y_p) \rangle, \quad \text{for all } Y_p \in M_p.$$

Consequently, $\mathsf{T}(\nabla'_{X_p} \xi)$ depends only on X_p and $\xi(p)$.

PROOF. If Y is a vector field tangent along M which extends Y_p , then $\langle \xi, Y \rangle = 0$, so

$$\begin{aligned} 0 &= X_p(\langle \xi, Y \rangle) = \langle \nabla'_{X_p} \xi, Y_p \rangle + \langle \xi(p), \nabla'_{X_p} Y \rangle \\ &= \langle \nabla'_{X_p} \xi, Y_p \rangle + \langle \xi(p), s(X_p, Y_p) \rangle, \end{aligned}$$

since $\xi(p) \in M_p^\perp$, by assumption. ♦

For any vector $\xi_p \in M_p^\perp$, we will define $A_{\xi_p}: M_p \rightarrow M_p$ as follows. For each $X_p \in M_p$, we let $A_{\xi_p}(X_p) \in M_p$ be the unique vector satisfying

$$\langle A_{\xi_p}(X_p), Y_p \rangle = \langle s(X_p, Y_p), \xi_p \rangle \quad \text{for all } Y_p \in M_p.$$

By Proposition 12, we also have

$$A_{\xi_p}(X_p) = -\mathsf{T}(\nabla'_{X_p} \xi),$$

where ξ is any normal vector field extending ξ_p . When M is a hypersurface in \mathbb{R}^{n+1} with unit normal vector field v , the map $A_{v_p}: M_p \rightarrow M_p$ is the same as $-dv: M_p \rightarrow M_p$.

For the normal component $\perp(\nabla'_{X_p} \xi)$ we will simply introduce a new symbol, just as we did for $\perp(\nabla'_{X_p} Y)$. For a section ξ of the normal bundle of M , and for $X_p \in M_p$, we define

$$D_{X_p} \xi = \perp(\nabla'_{X_p} \xi) \in M_p^\perp.$$

Unlike the case of $\perp(\nabla'_{X_p} Y)$, the value of $\perp(\nabla'_{X_p} \xi)$ depends on the values of ξ in a neighborhood of p , not just on $\xi(p)$.

13. PROPOSITION. The map $(X_p, \xi) \mapsto D_{X_p} \xi$ is a connection on the normal bundle $\text{Nor } M$; that is (compare pg. II.227 and also pg. II.346),

- (1) $D_{X_p+Y_p} \xi = D_{X_p} \xi + D_{Y_p} \xi$
- (2) $D_{X_p}(\xi + \eta) = D_{X_p} \xi + D_{X_p} \eta$
- (3) $D_{aX_p} \xi = a D_{X_p} \xi$ for all $a \in \mathbb{R}$
- (4) $D_{X_p} f \cdot \xi = f(p) \cdot D_{X_p} \xi + X_p(f) \cdot \xi(p)$ for all C^∞ functions f
- (5) If X is a C^∞ vector field and ξ is a C^∞ section of the normal bundle, then $p \mapsto D_{X_p} \xi$ is also C^∞ .

Moreover, D is compatible with the metric $\langle \cdot, \cdot \rangle$ on the normal bundle:

$$X_p(\langle \xi, \eta \rangle) = \langle D_{X_p} \xi, \eta \rangle + \langle \xi, D_{X_p} \eta \rangle.$$

PROOF. All properties follow immediately from the corresponding properties for ∇' . ♦

We will call D the **normal connection** for the imbedding $M \subset N$. With the notation we have just introduced, we may now write the decomposition $\nabla'_{X_p} \xi = \mathbf{T}(\nabla'_{X_p} \xi) + \mathbf{L}(\nabla'_{X_p} \xi)$ as

The Weingarten Equations: $\nabla'_{X_p} \xi = -A_{\xi_p}(X_p) + D_{X_p} \xi.$

In the case of a hypersurface, we used the second fundamental form s and a unit normal vector field v to define a real-valued second fundamental form \mathbf{II} . In the general case, we choose v_{n+1}, \dots, v_m to be everywhere orthonormal sections of E defined in a neighborhood of a point and we define $m - n$ real-valued **second fundamental forms** \mathbf{II}' by

$$\mathbf{II}'(X_p, Y_p) = \langle \nabla'_{X_p} Y_p, v_r(p) \rangle = \langle s(X_p, Y_p), v_r(p) \rangle \quad r = m + 1, \dots, n.$$

We thus have

$$s(X_p, Y_p) = \sum_r \mathbf{II}'(X_p, Y_p) \cdot v_r(p).$$

Notice that the set $\{\mathbf{II}'\}$ depends on the choice of the $\{v_r\}$; there are many possible choices, unlike the case of a hypersurface, where the choice of the single unit normal field v was essentially unique. Using the \mathbf{II}' instead of s , we can write

Gauss' Equation:

$$\begin{aligned} \langle R'(X, Y)Z, W \rangle &= \langle R(X, Y)Z, W \rangle \\ &+ \sum_r \{ \mathbf{II}'(X, Z) \mathbf{II}'(Y, W) - \mathbf{II}'(Y, Z) \mathbf{II}'(X, W) \}. \end{aligned}$$

Since the tensors \mathbf{II}' give us an explicit expression for s , they also essentially give us an expression for the $A_{v_r}(p)$, for the equation

$$\langle A_{v_r(p)}(X_p), Y_p \rangle = \langle v_r(p), s(X_p, Y_p) \rangle = \mathbf{II}'(X_p, Y_p)$$

determines $A_{v_r(p)}(X_p)$. We also want quantities by means of which we can express D . So we introduce certain 1-forms, the **normal fundamental forms** β_r^s , by

$$\beta_r^s(X_p) = \langle \nabla'_{X_p} v_r, v_s \rangle = \langle D_{X_p} v_r, v_s \rangle.$$

Then

$$D_{X_p} v_r = \sum_s \beta_r^s(X_p) \cdot v_s,$$

and $D_{X_p} \xi$ can be computed for any $\xi = \sum_r a^r v_r$ by using Proposition 13. Notice that since $\langle v_r, v_s \rangle = 1$ or 0 , we have $\beta_r^s = -\beta_s^r$. In particular, for hypersurfaces we have the single 1-form $\beta_{n+1}^{n+1} = 0$.

14. THEOREM. Let M be a submanifold of N , with corresponding s and D . Then for all vector fields X, Y, Z tangent along M we have

$$\begin{aligned} \perp(R'(X, Y)Z) &= [D_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z)] \\ &\quad - [D_Y(s(X, Z)) - s(\nabla_Y X, Z) - s(X, \nabla_Y Z)]. \end{aligned}$$

If ν_{n+1}, \dots, ν_m are everywhere orthonormal sections of $\text{Nor } M$, with corresponding Π^r and β_r^s , then for all vectors $X, Y, Z \in M_p$ we have

The Codazzi-Mainardi Equations:

$$\begin{aligned} \langle R'(X, Y)Z, \nu^r(p) \rangle &= (\nabla_X \Pi^r)(Y, Z) - (\nabla_Y \Pi^r)(X, Z) \\ &\quad + \sum_s \Pi^s(Y, Z) \beta_s^r(X) - \Pi^s(X, Z) \beta_s^r(Y). \end{aligned}$$

PROOF. The first equation is precisely equation (3) in the proof of Theorem I-11. Now Proposition 12 gives

$$\begin{aligned} (1) \quad D_X(s(Y, Z)) &= D_X\left(\sum_s \Pi^s(Y, Z) \nu_s\right) \\ &= \sum_s X(\Pi^s(Y, Z)) \cdot \nu_s + \sum_s \Pi^s(Y, Z) D_X \nu_s. \end{aligned}$$

Moreover,

$$(2) \quad s(\nabla_X Y, Z) + s(Y, \nabla_X Z) = \sum_s \Pi^s(\nabla_X Y, Z) \cdot \nu_s + \sum_s \Pi^s(Y, \nabla_X Z) \cdot \nu_s.$$

Then (1) and (2) give

$$\begin{aligned} (3) \quad &D_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z) \\ &= \sum_s [X(\Pi^s(Y, Z)) - \Pi^s(\nabla_X Y, Z) - \Pi^s(Y, \nabla_X Z)] \cdot \nu_s \\ &\quad + \sum_s \Pi^s(Y, Z) D_X \nu_s \\ &= \sum_s (\nabla_X \Pi^s)(Y, Z) \cdot \nu_s + \sum_s \Pi^s(Y, Z) D_X \nu_s \quad \text{by Corollary II.6-5*}. \end{aligned}$$

Hence

$$\begin{aligned} (4) \quad \langle D_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z), \nu_r \rangle \\ = (\nabla_X \Pi^r)(Y, Z) + \sum_s \Pi^s(Y, Z) \beta_s^r(X). \end{aligned}$$

*As on pg. III.11, we really need this Corollary for tensors of type $\binom{k}{0}$.

Naturally there is a similar equation with X and Y interchanged. Substituting into the first part of the Theorem, we obtain the Codazzi-Mainardi equations. ♦

The Codazzi-Mainardi equations which we have derived here are obviously a lot less satisfying than they were in the case of a hypersurface, since the set $\{\Pi'\}$ is not unique. As a matter of fact, the nicest form of the Codazzi-Mainardi equation is obtained by looking a little more closely at the expression

$$D_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z)$$

which appears in the first part of Theorem 14. A quick check shows that this expression is linear in X , Y , and Z over the C^∞ functions. So the value of this expression at p depends only on X_p, Y_p, Z_p . To obtain an explicit description of this function of three vectors, we consider s as a section of the bundle $\text{Hom}(TM \times TM, \text{Nor } M)$ whose fibre at p is the vector space of all bilinear maps $M_p \times M_p \rightarrow M_p^\perp$. Now, using the connection ∇ in TM and the connection D in $\text{Nor } M$, a connection $\tilde{\nabla}$ in the bundle $\text{Hom}(TM \times TM, \text{Nor } M)$ can be defined in the following natural way. Given

$$\begin{cases} \text{a section } \psi \text{ of } \text{Hom}(TM \times TM, \text{Nor } M), \\ \text{a vector } X_p \in M_p, \end{cases}$$

we want to have a bilinear map

$$\tilde{\nabla}_{X_p} \psi : M_p \times M_p \rightarrow M_p^\perp,$$

so we want to define

$$(\tilde{\nabla}_{X_p} \psi)(Y_p, Z_p), \quad \text{for } Y_p, Z_p \in M_p.$$

Let c be a curve in M with $c'(0) = X_p$, and let

$$\begin{aligned} \tau_h &= \text{the parallel translation in } TM \text{ along } c \\ &\quad \text{from } c(0) \text{ to } c(h) \text{ determined by } \nabla, \\ \rho_h &= \text{the parallel translation in } \text{Nor } M \text{ along } c \\ &\quad \text{from } c(0) \text{ to } c(h) \text{ determined by } D. \end{aligned}$$

Then we define

$$(\tilde{\nabla}_{X_p} \psi)(Y_p, Z_p) = \lim_{h \rightarrow 0} \frac{1}{h} [\rho_h^{-1}(\psi(c(h))(\tau_h Y_p, \tau_h Z_p)) - \psi(p)(Y_p, Z_p)].$$

[Notice that if $\text{Nor } M$ were just the trivial bundle $M \times \mathbb{R}$, making ψ essentially a tensor of type $\binom{2}{0}$, and D were the flat connection, with parallel translation the same along all curves, taking (p, a) to (q, a) for all $p, q \in M$ and $a \in \mathbb{R}$, then this definition would reduce to the definition of $\nabla_{X_p} \psi$ already given (pg. II.235). On the other hand, if $\text{Nor } M$ were TM , with the connection ∇ , then this definition would reduce to the definition of $\nabla_{X_p} \psi$ when ψ is a tensor of type $\binom{2}{1}$.] Now it is easy to see that if Y and Z are vector fields, then

$$(\tilde{\nabla}_{X_p} \psi)(Y_p, Z_p) = D_{X_p}(\psi(Y, Z)) - \psi(\nabla_{X_p} Y, Z_p) - \psi(Y_p, \nabla_{X_p} Z)$$

[Corollary II.6-5 is the special case when $\text{Nor } M$ is the trivial bundle]. We can therefore also express the Codazzi-Mainardi equations in an intrinsic form:

15. COROLLARY. Let M be a submanifold of N . Then for all vectors $X, Y, Z \in M_p$ we have

The Codazzi-Mainardi Equations:

$$\perp(R'(X, Y)Z) = (\tilde{\nabla}_X s)(Y, Z) - (\tilde{\nabla}_Y s)(X, Z)$$

where $\tilde{\nabla}$ is the covariant differentiation on $\text{Hom}(TM \times TM, \text{Nor } M)$ determined by the covariant differentiations ∇ on TM and D on the normal bundle $\text{Nor } M$.

The Gauss and Codazzi-Mainardi equations tell us what $\langle R'(X, Y)Z, W \rangle$ is when all four vectors are in M_p , or when three are in M_p and one is in M_p^\perp (which one doesn't matter, because the symmetry properties of R' allow us to express all possibilities in terms of the one where $W \in M_p^\perp$). We can just as well ask what $\langle R'(X, Y)Z, W \rangle$ is when two vectors are in M_p and two are in M_p^\perp . The answer is known (though not very well known) as the Ricci equations, or sometimes as the Ricci-Kühne equations. We need one more definition. Given $X, Y \in M_p$, and an orthonormal basis U_1, \dots, U_n of M_p , we set

$$\text{II}^r * \text{II}^s(X, Y) = \sum_{i=1}^n \text{II}^r(X, U_i) \cdot \text{II}^s(Y, U_i).$$

It is easy to check that $\text{II}^r * \text{II}^s$ does not depend on the choice of the orthonormal basis U_1, \dots, U_n . Classically, $\text{II}^r * \text{II}^s$ would be written as a contraction involving the components of II^r, II^s and the metric $\langle \cdot, \cdot \rangle^*$ on T^*M (compare pg. III.130).

16. THEOREM. Let M be a submanifold of N , with corresponding s , A , and D . Then for all vector fields X and Y tangent along M , and all sections ξ of the normal bundle $\text{Nor } M$ we have

$$\begin{aligned} \perp(R'(X, Y)\xi) &= s(A_\xi(X), Y) - s(A_\xi(Y), X) \\ &\quad + [D_X(D_Y\xi) - D_Y(D_X\xi) - D_{[X, Y]}\xi]. \end{aligned}$$

If v_{n+1}, \dots, v_m are everywhere orthonormal sections of $\text{Nor } M$, with corresponding Π^r and β_r^s , then for all vectors $X, Y \in M_p$ we have

The Ricci Equations:

$$\begin{aligned} \langle R'(X, Y)v_r(p), v_s(p) \rangle &= \Pi^r * \Pi^s(X, Y) - \Pi^r * \Pi^s(Y, X) \\ &\quad + (\nabla_X \beta_r^s)(Y) - (\nabla_Y \beta_r^s)(X) \\ &\quad + \sum_w \beta_w^s(X) \beta_r^w(Y) - \beta_w^s(Y) \beta_r^w(X). \end{aligned}$$

(Notice that these equations are trivial if M is a hypersurface.)

PROOF. The Weingarten equations and the Gauss formulas give

$$\begin{aligned} \nabla'_X(\nabla'_Y\xi) &= -\nabla'_X(A_\xi(Y)) + \nabla'_X(D_Y\xi) \\ &= -\nabla_X(A_\xi(Y)) - s(X, A_\xi(Y)) - A_{(D_Y\xi)}(X) + D_X(D_Y\xi), \end{aligned}$$

and hence

$$(1) \quad \perp(\nabla'_X(\nabla'_Y\xi)) = -s(X, A_\xi(Y)) + D_X(D_Y\xi),$$

$$(1') \quad \perp(\nabla'_Y(\nabla'_X\xi)) = -s(Y, A_\xi(X)) + D_Y(D_X\xi).$$

Also,

$$\nabla'_{[X, Y]}\xi = A_\xi([X, Y]) + D_{[X, Y]}\xi,$$

so

$$(2) \quad \perp(\nabla'_{[X, Y]}\xi) = D_{[X, Y]}\xi.$$

Equations (1), (1'), and (2) give the first part of the theorem.

Now if U_1, \dots, U_n is an orthonormal basis for M_p , then

$$\begin{aligned} A_{v_r(p)}(X_p) &= \sum_{i=1}^n \langle A_{v_r(p)}(X_p), U_i \rangle \cdot U_i \\ &= \sum_{i=1}^n \langle v_r(p), s(X_p, U_i) \rangle \cdot U_i \\ &= \sum_{i=1}^n \Pi^r(X_p, U_i) \cdot U_i, \end{aligned}$$

so

$$\begin{aligned}
 (3) \quad \langle s(A_{v_r(p)}(X_p), Y_p), v_s(p) \rangle &= \Pi^s(A_{v_r(p)}(X_p), Y_p) \\
 &= \Pi^s\left(\sum_{i=1}^n \Pi^r(X_p, U_i) \cdot U_i, Y_p\right) \\
 &= \sum_{i=1}^n \Pi^r(X_p, U_i) \cdot \Pi^s(Y_p, U_i) \\
 &= \Pi^r * \Pi^s(X_p, Y_p),
 \end{aligned}$$

$$(3') \quad \langle s(A_{v_r(p)}(Y_p), X_p), v_s(p) \rangle = \Pi^r * \Pi^s(Y_p, X_p).$$

We also have

$$\begin{aligned}
 D_X(D_Y v_r) &= \sum_w D_X(\beta_r^w(Y) \cdot v_w) \\
 &= \sum_w X(\beta_r^w(Y)) \cdot v_w + \sum_w \beta_r^w(Y) \cdot D_X v_w \\
 &= \sum_w X(\beta_r^w(Y)) \cdot v_w + \sum_w \beta_r^w(Y) \cdot \sum_v \beta_w^v(X) v_v,
 \end{aligned}$$

so

$$(4) \quad \langle D_X(D_Y v_r), v_s \rangle = X(\beta_r^s(Y)) + \sum_w \beta_w^s(X) \beta_r^w(Y),$$

$$(4') \quad \langle D_Y(D_X v_r), v_s \rangle = Y(\beta_r^s(X)) + \sum_w \beta_w^s(Y) \beta_r^w(X).$$

Also,

$$\begin{aligned}
 (5) \quad \langle D_{[X,Y]} v_r, v_s \rangle &= \langle D_{\nabla_X Y} v_r, v_s \rangle - \langle D_{\nabla_Y X} v_r, v_s \rangle \\
 &= \beta_r^s(\nabla_X Y) - \beta_r^s(\nabla_Y X).
 \end{aligned}$$

From (4), (4'), and (5) we get

$$\begin{aligned}
 (6) \quad &\langle D_X(D_Y v_r) - D_Y(D_X v_r) - D_{[X,Y]} v_r, v_s \rangle \\
 &= [X(\beta_r^s(Y)) - \beta_r^s(\nabla_X Y)] - [Y(\beta_r^s(X)) - \beta_r^s(\nabla_Y X)] \\
 &\quad + \sum_w \beta_w^s(X) \beta_r^w(Y) - \beta_w^s(Y) \beta_r^w(X) \\
 &= (\nabla_X \beta_r^s)(Y) - (\nabla_Y \beta_r^s)(X) + \sum_w \beta_w^s(Y) \beta_r^w(X) - \beta_w^s(X) \beta_r^w(Y) \\
 &\quad \text{by Corollary II.6-5.}
 \end{aligned}$$

Substituting (3), (3'), (6) into the first part of the theorem, we obtain the Ricci equations. ♦

The expression $D_X(D_Y\xi) - D_Y(D_X\xi) - D_{[X,Y]}\xi$ which appears in the first part of Theorem 16 can be treated just like the expressions which arose in Theorem 14. In fact, for any connection D in any vector bundle $\varpi: E \rightarrow M$, the map

$$(X, Y, \xi) \mapsto D_X(D_Y\xi) - D_Y(D_X\xi) - D_{[X,Y]}\xi$$

(for vector fields X, Y on M and sections ξ of E) is linear in X, Y , and ξ over the C^∞ functions. Consequently, its value at $p \in M$ depends only on $X_p, Y_p, \xi(p)$: we already know this for the two vector fields X, Y (Theorem I.4-2), and the proof for the section ξ is essentially the same. We therefore have a well-defined map

$$R_D = R_D(p): M_p \times M_p \times \varpi^{-1}(p) \rightarrow \varpi^{-1}(p),$$

the **curvature** of the connection D , given by

$$R_D(p)(X_p, Y_p)\xi_p = D_X(D_Y\xi)(p) - D_Y(D_X\xi)(p) - D_{[X,Y]}\xi(p),$$

for any vector fields X and Y extending X_p and Y_p , and any section ξ of E with $\xi(p) = \xi_p$. Thus we can state a more intrinsic form of the Ricci equations:

17. COROLLARY. Let M be a submanifold of N , with corresponding s and A . Then for all vectors $X, Y \in M_p$ and $\xi \in M_p^\perp$ we have

The Ricci Equations:

$$\perp(R'(X, Y)\xi) = R_D(X, Y)\xi + s(A_\xi(X), Y) - s(A_\xi(Y), X)$$

where R_D is the curvature of the connection D in $\text{Nor } M$.

The Gauss, Codazzi-Mainardi, and Ricci equations are the only general equations which we have for submanifolds of a Riemannian manifold. It would not be reasonable to expect an interesting formula for $\langle R'(X, Y)Z, W \rangle$ when *three* of the vectors are in M_p^\perp . For if $X, Y, Z \in M_p^\perp$, then $R'(X, Y, Z)$ has nothing to do with M at all, and $\langle R'(X, Y)Z, W \rangle$ would depend only on the position of M_p . The classical reason for resting content with these three equations was somewhat different. In Chapter 2 we saw (at least in a special case) that the Gauss and Codazzi-Mainardi equations were precisely the integrability conditions for the Gauss formulas. It turned out that the integrability conditions for the Weingarten equations reduced to the Codazzi-Mainardi equations, but this was only because we happened to be dealing with a hypersurface. In general, the integrability conditions for the Weingarten equations lead to two

sets of equations; one set reduces to the Codazzi-Mainardi equations, while the other set is precisely the Ricci equations [notice that our proof of Theorem 16 essentially investigated integrability conditions also, for we compared $\nabla'_X(\nabla'_Y\xi)$ with $\nabla'_Y(\nabla'_X\xi)$]. It was therefore clear to classical differential geometers that the Π^r and β_r^s determine an n -dimensional submanifold of \mathbb{R}^m up to Euclidean motion, and that any set of Π^r and β_r^s comes from some submanifold if the three fundamental equations are satisfied. In order to derive these results without writing everything out in very classical terms, we will first see what our fundamental equations say in terms of moving frames.

Consider an adapted orthonormal moving frame $X_1, \dots, X_n, X_{n+1}, \dots, X_m$ on M . As in Chapter 1, we let θ^i , ω_j^i , and Ω_j^i be the dual forms, connection forms, and curvature forms on M for the frame X_1, \dots, X_n , and we let ϕ^α , ψ_β^α , and Ψ_β^α be the forms on N for the frame X_1, \dots, X_m . Then on TM we have

$$\phi^i = \theta^i, \quad \phi^r = 0.$$

By looking at the first structural equations, we found that

$$\psi_j^i = \omega_j^i,$$

and that there are unique functions s_{ij}^r on M satisfying

$$\psi_j^r = \sum_i s_{ij}^r \theta^i, \quad s_{ij}^r = s_{ji}^r.$$

These functions are related to s by the equation (pg. III.19)

$$s(X_j, Y_k) = \sum_r s_{jk}^r X_r.$$

So if we choose the orthonormal vectors X_{n+1}, \dots, X_m to be our v_{n+1}, \dots, v_m , then the Π^r are given simply by

$$(a) \quad \Pi^r(X_r, X_k) = s_{kr}^r \implies \psi_j^r(X) = \Pi^r(X, X_j),$$

while the normal forms β_r^s are simply

$$(b) \quad \beta_r^s = \psi_r^s \quad \text{on } TM.$$

Since the map A is determined by

$$\begin{aligned} \langle A_{v_r}(X_i), X_j \rangle &= \langle v_r, s(X_i, X_j) \rangle \\ &= s_{ij}^r, \end{aligned}$$

we also have the explicit formula

$$A_{v_r}(X_i) = \sum_j s_{ij}^r X_j.$$

More important, we have

$$\begin{aligned} \text{(c)} \quad \Pi^r * \Pi^s(X_i, X_j) &= \sum_{k=1}^n \Pi^r(X_i, X_k) \Pi^s(X_j, X_k) \\ &= \sum_{k=1}^n s_{ik}^r s_{jk}^s. \end{aligned}$$

Now let us look at the second structural equation

$$d\psi_\beta^\alpha = - \sum_\gamma \psi_\lambda^\alpha \wedge \psi_\beta^\lambda + \Psi_\beta^\alpha.$$

If we restrict to TM , and choose various ranges for the indices, we obtain the following three equations (for the first we also use the structural equation on M , as in Chapter 1):

$$\text{(A)} \quad \Psi_j^i = \Omega_j^i - \sum_r \psi_i^r \wedge \psi_j^r$$

$$\text{(B)} \quad d\psi_j^r = - \sum_i \psi_i^r \wedge \omega_j^i - \sum_w \psi_w^r \wedge \psi_j^w + \Psi_j^r$$

$$\text{(C)} \quad d\psi_r^s = \sum_i \psi_i^s \wedge \psi_r^i - \sum_w \psi_w^s \wedge \psi_r^w + \Psi_r^s.$$

Using equation (a) we see immediately that equation (A) is precisely equivalent to Gauss' equation (in the form given on page 32). For equations (B) and (C) we recall that for a 1-form η we have (pg. I.215)

$$\text{(d)} \quad d\eta(X_k, X_l) = X_k(\eta(X_l)) - X_l(\eta(X_k)) - \eta([X_k, X_l]).$$

We also have

$$\text{(e)} \quad [X_k, X_l] = \nabla_{X_k} X_l - \nabla_{X_l} X_k = \sum_i \omega_l^i(X_k) X_i - \omega_k^i(X_l) X_i$$

and (Corollary II.6-5)

$$\text{(f)} \quad (\nabla_{X_k} \Pi^r)(X_l, X_j) = X_k(\Pi^r(X_l, X_j)) - \Pi^r(\nabla_{X_k} X_l, X_j) - \Pi^r(X_l, \nabla_{X_k} X_j)$$

$$\text{(g)} \quad (\nabla_{X_k} \beta_r^s)(X_l) = X_k(\beta_r^s(X_l)) - \beta_r^s(\nabla_{X_k} X_l).$$

When we apply equations (B) and (C) to (X_k, X_l) , and use equations (a)–(g), we find that (B) and (C) are equivalent to the Codazzi-Mainardi equations in Theorem 14, and the Ricci equations in Theorem 16, respectively. Equations (A), (B), (C), involving differential forms, are much more convenient for considering questions connected with integrability conditions.

18. THEOREM. (1) Let $M, \bar{M} \subset \mathbb{R}^m$ be two connected n -manifolds imbedded in \mathbb{R}^m , let v_{n+1}, \dots, v_m be everywhere orthonormal sections for the normal bundle of M , and let $\bar{v}_{n+1}, \dots, \bar{v}_m$ be everywhere orthonormal sections for the normal bundle of \bar{M} . Let I, II^r, β_r^s be the first, second, and normal fundamental forms for M (defined with respect to the $\{v^r\}$), and define $\bar{I}, \bar{II}^r, \bar{\beta}_r^s$ similarly. Let $\phi: M \rightarrow \bar{M}$ be a diffeomorphism which preserves all the fundamental forms:

$$\phi^* \bar{I} = I, \quad \phi^* \bar{II}^r = II^r, \quad \phi^* \bar{\beta}_r^s = \beta_r^s.$$

Then there is a Euclidean motion A such that $\phi = A|M$ and $A_*(v_r) = \bar{v}_r$ for $r = n+1, \dots, m$.

(2) Let $(M, \langle \cdot, \cdot \rangle)$ be an n -dimensional Riemannian manifold with curvature tensor R . For $r, s = n+1, \dots, m$, let S^r be symmetric tensors on M , covariant of order 2, and let b_r^s be 1-forms on M with $b_r^s = -b_s^r$. Suppose that the S^r and b_r^s satisfy

(1) Gauss' Equation:

$$0 = \langle R(X, Y)Z, W \rangle - \sum_r S^r(Y, Z)S^r(X, W) + S^r(X, Z)S^r(Y, W)$$

(2) The Codazzi-Mainardi Equations:

$$\begin{aligned} 0 &= (\nabla_X S^r)(Y, Z) - (\nabla_Y S^r)(X, Z) \\ &\quad + \sum_s \{S^s(Y, Z)b_s^r(X) - S^s(X, Z)b_s^r(Y)\} \end{aligned}$$

(3) The Ricci Equations:

$$\begin{aligned} 0 &= S^r * S^s(X, Y) - S^r * S^s(Y, X) + (\nabla_X b_r^s)(Y) - (\nabla_Y b_r^s)(X) \\ &\quad + \sum_w \{b_w^s(X)b_r^w(Y) - b_w^s(Y)b_r^w(X)\}. \end{aligned}$$

Then for every point of M there is a neighborhood U and an isometric imbedding $f: U \rightarrow \mathbb{R}^m$ such that there are everywhere orthonormal sections v_{n+1}, \dots, v_m of the normal bundle of $f(U)$ in \mathbb{R}^m for which the corresponding forms II^r and β_r^s on $f(U)$ satisfy

$$s^r = f^* II^r, \quad b_r^s = f^* \beta_r^s.$$

PROOF. We will consider the proof of (2) first, since the proof of (1) will come along for free. Since we are trying to prove a local result, we might as well

assume that M is \mathbb{R}^n . Let X_1, \dots, X_n be an orthonormal moving frame for $\langle \cdot, \cdot \rangle$ on \mathbb{R}^n , with dual forms θ^i , and connection forms ω_j^i . Define 1-forms ψ_β^α ($1 \leq \alpha, \beta \leq m$) on \mathbb{R}^n as follows:

$$\begin{aligned}\psi_j^i &= \omega_j^i & 1 \leq i, j \leq n \\ \psi_j^r(X) &= S^r(X, X_j) & 1 \leq j \leq n < r \leq m \\ \psi_s^r &= b_s^r & n < r, s \leq m.\end{aligned}$$

Then the forms ψ_β^α satisfy two crucial equations:

$$(*) \quad \sum_{j=1}^n \psi_j^r \wedge \theta^j = 0 \quad r = n+1, \dots, m$$

$$(**) \quad d\psi_\beta^\alpha = - \sum_{\gamma=1}^m \psi_\gamma^\alpha \wedge \psi_\beta^\gamma \quad \alpha, \beta = 1, \dots, m.$$

Equation (*) follows directly from the definition of ψ_j^r and symmetry of S^r . Equation (**) follows from the Gauss, Codazzi-Mainardi, and Ricci equations in the hypothesis. This should be clear from our verifications, prior to the statement of the theorem, that equations (A), (B), and (C) are equivalent to Gauss' equation on page 32, and to the Codazzi-Mainardi equations in Theorem 14 and 16, respectively.

Now suppose for the moment that we have an immersion $f: (\mathbb{R}^n, \langle \cdot, \cdot \rangle) \rightarrow \mathbb{R}^m$, and orthonormal sections v_{n+1}, \dots, v_m of the normal bundle of $f(\mathbb{R}^n)$. Identifying tangent vectors of \mathbb{R}^m with elements of \mathbb{R}^m , as usual, we thus have a map

$$v = (v_1, \dots, v_m) = (f_*(X_1), \dots, f_*(X_n), v_{n+1}, \dots, v_m): \mathbb{R}^n \rightarrow \mathbb{R}^{m^2}.$$

If f is an isometry and $S^r = f^* \Pi^r$ and $b_r^s = f^* \beta_r^s$, then the components v_α^β of the functions v_α will satisfy

$$(1) \quad dv_\alpha^\beta = \sum_{\gamma=1}^m v_\gamma^\beta \cdot \psi_\alpha^\gamma.$$

So we will first show that a map $v = (v_1, \dots, v_m): \mathbb{R}^n \rightarrow \mathbb{R}^{m^2}$ satisfying (1) can be found. The idea of the proof is to look for the graph $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^{m^2}$ of v (compare pg. II.264). Let $\pi_1: \mathbb{R}^n \times \mathbb{R}^{m^2} \rightarrow \mathbb{R}^n$ and $\pi_2: \mathbb{R}^n \times \mathbb{R}^{m^2} \rightarrow \mathbb{R}^{m^2}$ be the projections on the first and second factors, and let $\{x_\alpha^\beta\}$ be the standard coordinate system on \mathbb{R}^{m^2} . It is easy to see that if $v: \mathbb{R}^n \rightarrow \mathbb{R}^{m^2}$ satisfying (1)

exists, then its graph $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^{m^2}$ is a submanifold on which the m^2 linearly independent 1-forms

$$(2) \quad d(x_\alpha^\beta \circ \pi_2) - \sum_{\gamma=1}^m (x_\gamma^\beta \circ \pi_2) \cdot \pi_1^* \psi_\alpha^\gamma$$

all vanish. Conversely, if all these 1-forms vanish on an n -dimensional manifold $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^{m^2}$, then Γ is the graph of the desired function v . So by the Frobenius integrability theorem (I.7-14), we just have to show that the exterior derivative of each form (2) is in the ideal \mathcal{I} generated by these forms. Now

$$\begin{aligned} d\left(\sum_{\gamma=1}^m (x_\gamma^\beta \circ \pi_2) \cdot \pi_1^* \psi_\alpha^\gamma\right) &= \sum_{\gamma=1}^m d(x_\gamma^\beta \circ \pi_2) \wedge \pi_1^* \psi_\alpha^\gamma \\ &\quad + \sum_{\lambda=1}^m (x_\lambda^\beta \circ \pi_2) \cdot \pi_1^* d\psi_\alpha^\lambda \\ &= \sum_{\gamma=1}^m d(x_\alpha^\beta \circ \pi_2) \wedge \pi_1^* \psi_\alpha^\gamma \\ &\quad - \sum_{\gamma=1}^m (x_\gamma^\beta \circ \pi_2) \cdot \left(\sum_{\lambda=1}^m \pi_1^* \psi_\gamma^\lambda \wedge \pi_1^* \psi_\alpha^\gamma\right) \text{ by } (**) \\ &= \sum_{\gamma=1}^m \left(d(x_\alpha^\beta \circ \pi_2) - \sum_{\lambda=1}^m (x_\lambda^\beta \circ \pi_2) \cdot \pi_1^* \psi_\gamma^\lambda\right) \wedge \pi_1^* \psi_\alpha^\gamma, \end{aligned}$$

which is indeed in the ideal \mathcal{I} generated by the forms (2). Thus we see that there is a function $v: \mathbb{R}^n \rightarrow \mathbb{R}^{m^2}$ satisfying (I). In fact, we can choose $v(0)$ to be any linearly independent set of vectors in \mathbb{R}^{m^2} (just choose the integral submanifold of $\mathcal{I} = 0$ which passes through this set of vectors); in particular, we can choose $v(0)$ to be a set of orthonormal vectors.

We next note that, for the ordinary inner product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^m , the functions v_1, \dots, v_m satisfy

$$\begin{aligned} d(\langle v_\alpha, v_\beta \rangle) &= \langle dv_\alpha, v_\beta \rangle + \langle v_\alpha, dv_\beta \rangle \\ &= \sum_{\gamma=1}^m v_\beta^\gamma \cdot dv_\alpha^\gamma + \sum_{\gamma=1}^m v_\alpha^\gamma \cdot dv_\beta^\gamma \\ &= \sum_{\gamma, \lambda=1}^m v_\beta^\gamma v_\lambda^\gamma \psi_\alpha^\lambda + \sum_{\gamma, \lambda=1}^m v_\alpha^\gamma v_\lambda^\gamma \psi_\beta^\lambda \quad \text{by (1)} \\ &= \sum_{\lambda=1}^m \langle v_\beta, v_\lambda \rangle \psi_\alpha^\lambda + \sum_{\lambda=1}^m \langle v_\alpha, v_\lambda \rangle \psi_\beta^\lambda. \end{aligned}$$

In particular, for any curve c in \mathbb{R}^m , the functions

$$f_{\alpha\beta}(t) = \langle v_\alpha(c(t)), v_\beta(c(t)) \rangle$$

satisfy the differential equation

$$f_{\alpha\beta}'(t) = \sum_{\lambda=1}^m \psi_\alpha^\lambda(c'(t)) f_{\beta\lambda}(t) + \sum_{\lambda=1}^m \psi_\beta^\lambda(c'(t)) f_{\alpha\lambda}(t).$$

Since $\psi_\beta^\alpha = -\psi_\alpha^\beta$, this same equation is satisfied by the functions $f_{\alpha\beta}(t) = \delta_{\alpha\beta}$. So by uniqueness of solutions with a given initial condition, we conclude that v_1, \dots, v_m are orthonormal everywhere.

Now we want to show that there is actually a function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$f_*(X_i) = v_i \quad i = 1, \dots, n.$$

For the component functions f^1, \dots, f^m of f we want

$$df^\alpha(X_i(p)) = v_i^\alpha(p) \quad \text{i.e.,} \quad df^\alpha(p) = \sum_{j=1}^n v_j^\alpha(p) \cdot \theta^j(p).$$

To prove that f exists, we look for its graph $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^m$. We let $\pi_1: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $\pi_2: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ be the projections on the first and second factors, and we let $\{x^\alpha\}$ be the standard coordinate system on \mathbb{R}^m . The Γ we are seeking is a submanifold of $\mathbb{R}^n \times \mathbb{R}^m$ on which the 1-forms

$$(3) \quad d(x^\alpha \circ \pi_2) - \sum_{j=1}^n (v_j^\alpha \circ \pi_1) \cdot \pi_1^* \theta^j$$

all vanish. Now

$$\begin{aligned} d\left(\sum_{j=1}^n (v_j^\alpha \circ \pi_1) \cdot \pi_1^* \theta^j\right) &= \sum_{j=1}^n \pi_1^* dv_j^\alpha \wedge \pi_1^* \theta^j \\ &\quad - \sum_{i=1}^n (v_i^\alpha \circ \pi_1) \cdot \sum_{j=1}^n \pi_1^* \omega_j^i \wedge \pi_1^* \theta^j \\ &= \sum_{j=1}^n \sum_{\gamma=1}^m (v_\gamma^\alpha \circ \pi_1) \cdot \pi_1^* \psi_j^\gamma \wedge \pi_1^* \theta^j \\ &\quad - \sum_{i,j=1}^n (v_i^\alpha \circ \pi_1) \cdot \pi_1^* \omega_j^i \wedge \pi_1^* \theta^j \quad \text{by (1)} \\ &= \sum_{j=1}^n \sum_{r=n+1}^m (v_r^\alpha \circ \pi_1) \cdot \pi_1^* \psi_j^r \wedge \pi_1^* \theta^j \\ &= 0 \quad \text{by (*).} \end{aligned}$$

So the Frobenius integrability theorem proves that there is $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $f_*(X_i) = v_i$ ($i = 1, \dots, n$). Then f is an isometry, since X_1, \dots, X_n and v_1, \dots, v_n are both orthonormal sets. This proves part (2).

As for part (1), we first note that the required Euclidean motion A , if it exists, is unique [for if $p \in M$, then $A_*|M_p$ must be ϕ_{*p} , and $A_*(v_r(p))$ must be $\bar{v}_r(p)$]. Therefore it suffices to prove existence of A locally. By the usual argument, it suffices to show that $M = \bar{M}$ if for some $p \in M$ we have $p = \phi(p)$ and $M_p = \bar{M}_p$, and $\phi_{*p} = \text{identity}$ and $v_r(p) = \bar{v}_r$. But this follows immediately from the fact that there is just one function $v: \mathbb{R}^n \rightarrow \mathbb{R}^{m^2}$ satisfying (1), with a given value of $v(0)$. ♦

This classical formulation of Theorem 18 was given in order to emphasize the extra problems which arise when the codimension is greater than 1. But it is a very unsatisfactory way of bringing the normal bundle into the picture. After all, everywhere orthonormal sections v_{m+1}, \dots, v_n of the normal bundle can usually be found only in a neighborhood of each point. So the first part of Theorem 18 really makes sense only locally, even though it is supposed to be global. We have similar problems in the second part, where we would like to obtain a global result when M is simply connected. These defects are easily rectified, since we also have invariant statements of our fundamental equations. We need one simple bit of terminology. Let M and \bar{M} be C^∞ manifolds, and let $\varpi: E \rightarrow M$ and $\bar{\varpi}: \bar{E} \rightarrow \bar{M}$ be C^∞ vector bundles of the same dimension over M and \bar{M} , respectively. If $\phi: M \rightarrow \bar{M}$ is a diffeomorphism, then a C^∞ map $\tilde{\phi}: E \rightarrow \bar{E}$ is called a **bundle isomorphism covering ϕ** if:

(1) the diagram

$$\begin{array}{ccc} E & \xrightarrow{\tilde{\phi}} & \bar{E} \\ \varpi \downarrow & & \downarrow \bar{\varpi} \\ M & \xrightarrow{\phi} & \bar{M} \end{array}$$

commutes, so that $\tilde{\phi}$ takes $\varpi^{-1}(p)$ to $\bar{\varpi}^{-1}(\phi(p))$,

(2) each map

$$\tilde{\phi}|_{\varpi^{-1}(p)}: \varpi^{-1}(p) \rightarrow \bar{\varpi}^{-1}(\phi(p))$$

is a vector space isomorphism.

It then follows easily that $\tilde{\phi}$ is a diffeomorphism. Notice that if ξ is a section of E , then we have a section $\tilde{\phi}(\xi)$ of \bar{E} defined by

$$\tilde{\phi}(\xi)(q) = \tilde{\phi}(\xi(\phi^{-1}(q))) \quad q \in \bar{M}.$$

19. THEOREM. (1) Let M be a connected submanifold of \mathbb{R}^m with normal bundle $\varpi: \text{Nor } M \rightarrow M$, and corresponding second fundamental form s and normal connection D . Similarly, let \bar{M} be a connected submanifold with normal bundle $\bar{\varpi}: \text{Nor } \bar{M} \rightarrow \bar{M}$ and corresponding \bar{s} and \bar{D} . Let $\phi: M \rightarrow \bar{M}$ be an isometry. Suppose that there is a bundle isomorphism $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ covering ϕ such that $\tilde{\phi}$ preserves inner products, second fundamental forms, and normal connections:

$$\begin{aligned} \langle \tilde{\phi}(\xi), \tilde{\phi}(\eta) \rangle &= \langle \xi, \eta \rangle & \text{for all } \xi, \eta \in M_p^\perp \\ \tilde{\phi}(s(X, Y)) &= \bar{s}(\phi_* X, \phi_* Y) & \text{for all } X, Y \in M_p \\ \tilde{\phi}(D_X \xi) &= \bar{D}_{\phi_* X}(\tilde{\phi}(\xi)) & \text{for all } X \in M_p \text{ and all sections } \xi \text{ of } \text{Nor } M. \end{aligned}$$

Then there is a Euclidean motion A such that $\phi = A|_M$ and $\tilde{\phi} = A_*|_{\text{Nor } M}$.

(2) Let $(M, \langle \cdot, \cdot \rangle)$ be a simply-connected n -dimensional Riemannian manifold, with covariant differentiation ∇ and curvature tensor R . Let $\varpi: E \rightarrow M$ be an $(m-n)$ -dimensional vector bundle over M with a Riemannian metric $\{ \cdot, \cdot \}$, let δ be a connection on E compatible with $\{ \cdot, \cdot \}$, with curvature tensor R_δ , and let σ be a symmetric section of the bundle $\text{Hom}(TM \times TM, E)$. Denote by $\tilde{\nabla}$ the connection on $\text{Hom}(TM \times TM, E)$ determined by ∇ and δ ; and for $X \in M_p$ and $\xi \in \varpi^{-1}(p)$, let $A_\xi(X) \in M_p$ be the unique vector satisfying

$$\langle A_\xi(X), Y \rangle = \{ \xi, \sigma(X, Y) \} \quad \text{for all } Y \in M_p.$$

Suppose that σ and δ satisfy

(1) Gauss' Equation:

$$\langle R(X, Y)Z, W \rangle = \{ \sigma(Y, Z), \sigma(X, W) \} - \{ \sigma(X, Z), \sigma(Y, W) \}$$

(2) The Codazzi-Mainardi equations:

$$(\tilde{\nabla}_X \sigma)(Y, Z) = (\tilde{\nabla}_Y \sigma)(X, Z)$$

(3) The Ricci equations:

$$R_\delta(X, Y)\xi = \sigma(A_\xi(Y), X) - \sigma(A_\xi(X), Y).$$

Then there is an isometric immersion $f: M \rightarrow \mathbb{R}^m$ and a bundle isomorphism $\tilde{f}: E \rightarrow \{\text{normal bundle of } f(M)\}$ covering f such that

$$\begin{aligned} \langle \tilde{f}(\xi), \tilde{f}(\eta) \rangle &= \langle \xi, \eta \rangle & \text{for all } \xi, \eta \in \varpi^{-1}(p) \\ \tilde{f}(\sigma(X, Y)) &= s(f_* X, f_* Y) & \text{for all } X, Y \in M_p \\ \tilde{f}(\delta_X \xi) &= D_{\phi_* X}(\tilde{f}(\xi)) & \text{for all } X \in M_p \text{ and all sections } \xi \text{ of } E, \end{aligned}$$

where s and D are the second fundamental form and the normal connection for $f(M)$.

PROOF. It's just one big translation job locally. Then simple-connectivity is used to prove the global result, as in Problem 2-3. ♦

Naturally, Theorem 19 simplifies considerably for the case of hypersurfaces, when $\text{Nor } M$ is 1-dimensional. In part (1) we can dispense with the normal connection D , and in part (2) we can dispense with δ (and the Ricci equations). When we deal with *oriented* hypersurfaces, we can ignore $\text{Nor } M$ completely, since the orientation of M determines a unit normal field v , and thus a second fundamental form Π . In part (1), we simply need that ϕ is an isometry with $\phi^*\bar{\Pi} = \Pi$; then the Euclidean motion A of the conclusion is actually a proper Euclidean motion with $A_*v = \bar{v}$. In part (2), we simply supply our Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ with a symmetric tensor S , covariant of order 2, satisfying

$$\begin{aligned}\langle R(X, Y)Z, W \rangle &= S(Y, Z) \cdot S(X, W) - S(X, Z) \cdot S(Y, W) \\ (\nabla_X S)(Y, Z) &= (\nabla_Y S)(X, Z).\end{aligned}$$

In the general case, Theorem 19 is much less satisfying, for it does *not* tell us when a map $\phi: M \rightarrow \bar{M}$ between two submanifolds of \mathbb{R}^m is the restriction of a Euclidean motion; it only gives us information about maps $\phi: M \rightarrow \bar{M}$ together with bundle isomorphisms $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ covering ϕ (as a slight compensation, we find a Euclidean motion A preserving this additional structure). In the theory of curves we have a situation more closely resembling the case of hypersurfaces, since certain functions $\kappa_1, \dots, \kappa_{m-1}$ determine (“general”) curves parameterized by arclength. For curves in \mathbb{R}^m , the results of part B are actually a special case of Theorem 19, for the vector fields v_2, \dots, v_m along c give a trivialization of the normal bundle of c ; it is not hard to see (Problem 12) that if we choose $v_r = \mathbf{v}_r$, then the corresponding Π' and β_r^s are all expressible in terms of $\kappa_1, \dots, \kappa_{m-1}$. For higher dimensional submanifolds of higher codimension there is also a theory which determines “general” submanifolds, up to Euclidean motions, by means of tensors on the submanifold. Although this theory is certainly more appealing geometrically, it is rather elaborate, and is presented separately in Addendum 4.

For the present, we merely wish to generalize Theorem 19 by replacing \mathbb{R}^m with a more general ambient space. Now the first part of Theorem 19 simply isn't true for an arbitrary ambient space $(N, \langle \cdot, \cdot \rangle)$, which may not have any isometries onto itself except the identity. For example, we can easily construct a metric on \mathbb{R}^{n+1} such that the hypersurfaces $\mathbb{R}^n \times \{0\}$ and $\mathbb{R}^n \times \{1\}$ are isometric and have vanishing second fundamental forms, without there being any non-trivial isometry of \mathbb{R}^{n+1} onto itself, and in particular none taking $\mathbb{R}^n \times \{0\}$ to

$\mathbb{R}^n \times \{1\}$. It should be mentioned, however, that one can at least prove the following (Problem 12):

Suppose that M and \bar{M} are connected submanifolds of $(N, \langle \cdot, \cdot \rangle)$ and that there is a point $p \in M \cap \bar{M}$ with $M_p = \bar{M}_p$. Let $\phi: M \rightarrow \bar{M}$ be an isometry with $\phi(p) = p$ and $\phi_{*p}: M_p \rightarrow \bar{M}_p$ the identity. Suppose that there is a bundle isomorphism $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ covering ϕ which preserves inner products, second fundamental forms, and normal connections, and such that $\tilde{\phi}$ is the identity map on M_p^\perp . Then $M = \bar{M}$ and ϕ is the identity.

We encounter difficulties of another sort when we try to generalize the second part of Theorem 19 for an arbitrary ambient manifold $(N, \langle \cdot, \cdot \rangle)$. Now we don't even know what conditions to place on δ and σ , since the Codazzi-Mainardi and Ricci equations for δ and σ involve terms $R'(X, Y)Z$ which we cannot evaluate unless we already have the imbedding of M into N .

These difficulties do not arise when $(N, \langle \cdot, \cdot \rangle)$ is a complete simply-connected manifold of constant curvature K_0 . The Euclidean motions of part (I) will be replaced by the isometries $A: N \rightarrow N$; such isometries can be found taking any orthonormal frame at one point of N to any orthonormal frame at any other point (Problem 1-5). Moreover, the Codazzi-Mainardi and Ricci equations for a submanifold $M \subset N$ are exactly the same as in the Euclidean case, since

$$R'(X, Y)Z = K_0[\langle Y, Z \rangle X - \langle X, Z \rangle Y] \quad (\text{pg. III.11})$$

$$\Downarrow$$

$$\perp(R'(X, Y)Z) = 0 \quad \text{for } X, Y, Z \text{ tangent to } M$$

$$R'(X, Y)\xi = 0 \quad \text{for } X, Y, Z \text{ tangent to } M \text{ and } \xi \text{ normal to } M.$$

Gauss' equation, on the other hand, becomes (Corollary 1-12)

$$\begin{aligned} K_0[\langle X, W \rangle \cdot \langle Y, Z \rangle - \langle X, Z \rangle \cdot \langle Y, W \rangle] \\ = \langle R(X, Y)Z, W \rangle + \langle s(X, Z), s(Y, W) \rangle - \langle s(Y, Z), s(X, W) \rangle. \end{aligned}$$

For an adapted orthonormal moving frame on M , equations (B) and (C) have $\Psi_j^r = \Psi_r^s = 0$, while equation (A) becomes

$$(A') \quad K_0[\theta^i \wedge \theta^j] = \Omega_j^i - \sum_r \psi_i^r \wedge \psi_j^r,$$

since

$$\begin{aligned} \Psi_j^i(X, Y) &= \langle R'(X, Y)X_j, X_i \rangle = K_0[\langle X, X_i \rangle \cdot \langle Y, X_j \rangle - \langle X, X_j \rangle \cdot \langle Y, X_i \rangle] \\ &= K_0[\theta^i(X)\theta^j(Y) - \theta^i(Y)\theta^j(X)]. \end{aligned}$$

One fairly straightforward method of generalizing Theorem 19 is to regard N as $S^m(K_0) \subset \mathbb{R}^{m+1}$, or as $H^m(K_0) \subset \mathbb{R}^{m+1}$ with the Lorentzian metric. The details of this pleasant proof are left to Problems 13 and 14, partly because it uses a result from the next section, but mainly because we want to provide a (rather unpleasant) proof which involves only the intrinsic description of N as a manifold of constant curvature K_0 . The proof of Theorem 19 itself will not generalize at all, because it involves the natural identification of $T\mathbb{R}^m$ with $\mathbb{R}^m \times \mathbb{R}^m$, an identification that essentially depends on the fact that \mathbb{R}^m is flat. For a general manifold $(N, \langle \cdot, \cdot \rangle)$ of constant curvature, we will have to consider the tangent bundle TN , and work with Ehresmann connections.

Consider a principal bundle $\pi: P \rightarrow M$ with group G . For each $X \in \mathfrak{g} = \text{Lie algebra of } G$, we have defined (pg. II.311) the **fundamental vector field** on P corresponding to X ; we will change notation slightly and denote this vector field by $\sigma(X)$ [so that σ can still be used as in the statement of Theorem 19]. Recall that an Ehresmann connection on P is a \mathfrak{g} -valued 1-form on P . We will use a bold face Greek letter, like ω , for such connections. Here our aim is to avoid confusing ω with the (closely related) connection forms ω_j^i of a moving frame on a manifold. Recall that a **frame** u for M_p is an ordered basis $u = (u_1, \dots, u_n)$ of M_p , and that we have a principal bundle $F(TM) \rightarrow M$ with group $\text{GL}(n, \mathbb{R})$, where $F(TM)$ is the set of all frames at all $p \in M$. An Ehresmann connection $\omega = (\omega_j^i)$ on $F(TM)$ is a $\mathfrak{gl}(n, \mathbb{R})$ -valued 1-form on $F(TM)$. For any moving frame $s = (X_1, \dots, X_n): U \rightarrow F(TM)$ on an open set $U \subset M$ we thus have the matrix of 1-forms $\omega = s^*\omega$, and the assignment of $\omega = s^*\omega$ to $s = (X_1, \dots, X_n)$ is a (Cartan) connection on M ; then by defining $\nabla_X X_j = \sum_i \omega_j^i(X) X_i$, we obtain a covariant differentiation operator ∇ on M . Conversely, every Cartan connection on M comes from a unique Ehresmann connection ω in this way, and thus every covariant differentiation operator ∇ comes from a unique ω . More generally, given a k -dimensional vector bundle $\varpi: E \rightarrow M$, we let $F(E) \rightarrow M$ denote the set of all ordered bases u of $\varpi^{-1}(p)$, for all $p \in M$. Then $F(E) \rightarrow M$ is a principal bundle with group $\text{GL}(k, \mathbb{R})$, and Ehresmann connections ω on $F(E)$ correspond to covariant differentiation operators $(X, \xi) \mapsto \nabla_X \xi$ on sections ξ of E . In the special case of $F(TM)$ we also have the \mathbb{R}^n -valued **dual form** $\theta = (\theta^1, \dots, \theta^n)$; for any moving frame $s = (X_1, \dots, X_n)$, the forms $\theta^i = s^*\theta^i$ are just the dual forms for this moving frame.

When we have a vector bundle $\varpi: E \rightarrow M$ which has a Riemannian metric $\langle \cdot, \cdot \rangle$, it is often more convenient to consider the bundle $O(E) \subset F(E)$ consisting of *orthonormal* frames. Notice that for $X \in \mathfrak{o}(k) \subset \mathfrak{gl}(k, \mathbb{R})$, the vector field $\sigma(X)$ on $O(E)$ is just the restriction of the vector field $\sigma(X)$ which is defined on

$F(E)$. Now suppose we have a covariant differentiation $(X, \xi) \mapsto \nabla_X \xi$ which is compatible with the metric $\{ , \}$, so that

$$X(\{\xi, \eta\}) = \{\nabla_X \xi, \eta\} + \{\xi, \nabla_X \eta\}.$$

If ξ_1, \dots, ξ_k are local sections of E with $\{\xi_i, \xi_j\} = \delta_{ij}$, and we define the 1-forms ω_j^i ($1 \leq i, j \leq k$) by

$$\nabla_X \xi_j = \sum_{i=1}^k \omega_j^i(X) \cdot \xi_i,$$

then we have

$$(I) \quad \omega_j^i = -\omega_i^j.$$

Let $\omega = (\omega_j^i)$ be the Ehresmann connection on $F(E)$ corresponding to the covariant differentiation ∇ . We claim that $\omega|_{O(E)}$ actually takes values in $\mathfrak{o}(k) = \{\text{skew-symmetric } k \times k \text{ matrices}\}$. In fact, if Y is a vertical vector at some $u \in O(E)$, then $Y = \sigma(X)(u)$ for some $X \in \mathfrak{o}(k)$, and

$$\omega(Y) = \omega(\sigma(X)(u)) = X \in \mathfrak{o}(k).$$

On the other hand, every non-vertical vector at a frame $u \in O(E)$ at $p \in M$ is of the form $s_*(Z)$ for some local orthonormal section $s = (\xi_1, \dots, \xi_k)$ and some tangent vector $Z \in M_p$, and then

$$\omega(s_*(Z)) = s^* \omega(Z) = \omega(Z),$$

so the claim follows from equation (I). Thus we see that $\omega|_{O(E)}$ is an Ehresmann connection on the principal bundle $O(E)$. [Conversely, an Ehresmann connection on $O(E)$ clearly extends in a natural way to an Ehresmann connection on $F(E)$ whose corresponding ∇ is compatible with the metric $\{ , \}$.] It is clear that at any point $u \in O(E)$, the horizontal subspace for the connection $\omega|_{O(E)}$ is exactly the same as the horizontal subspace at $u \in F(E)$ for the connection ω . So for $Y_1, Y_2 \in O(E)_u$, the covariant differential $D(\omega|_{O(E)})$ has value

$$\begin{aligned} D(\omega|_{O(E)})(Y_1, Y_2) &= d(\omega|_{O(E)})(hY_1, hY_2) \\ &= d\omega(hY_1, hY_2) \\ &= \Omega(Y_1, Y_2). \end{aligned} \quad \left\{ \begin{array}{l} hY_i = \text{horizontal} \\ \text{component of } Y_i \\ \text{in either } F(E) \\ \text{or } O(E) \end{array} \right.$$

In other words, the curvature form of $\omega|_{O(E)}$ is just the restriction to $O(E)$ of the curvature form Ω of ω . So no confusion will arise if we simply use ω for the restriction $\omega|_{O(E)}$, and Ω for the curvature form of $\omega|_{O(E)}$.

We will apply these considerations, in particular, to the case where $E = TM$ is the tangent bundle of a Riemannian manifold $(M^n, \langle \cdot, \cdot \rangle)$ and ω is the Ehresmann connection corresponding to the Levi-Civita connection for $\langle \cdot, \cdot \rangle$. Thus we have $\mathfrak{o}(n)$ -valued forms $\theta^i, \omega_j^i, \Omega_j^i$ on $O(TM)$. For another Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$ we have, similarly, $\mathfrak{o}(m)$ -valued forms $\phi^\alpha, \psi_\beta^\alpha, \Psi_\beta^\alpha$ on $O(TN)$.

Now suppose that we are given a Riemannian manifold $(M^n, \langle \cdot, \cdot \rangle)$, and an $(m - n)$ -dimensional vector bundle $\varpi: E \rightarrow M$ with a Riemannian metric $\{ \cdot, \cdot \}$. Let $O(TM, E)$ denote the set of all pairs (u, v) where $u \in O(TM)$ and $v \in O(E)$ lie over the same point $p \in M$. Then $O(TM, E)$ is a principal bundle, whose group G is the set of all $m \times m$ matrices of the form

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \quad A \in O(n) \quad \text{and} \quad B \in O(m - n).$$

These bundles will come into our proof of the generalization of Theorem 19 in the following way. If we succeed in finding an isometric immersion $f: M \rightarrow N$, covered by an inner product preserving bundle isomorphism $\tilde{f}: E \rightarrow \{\text{normal bundle of } f(M)\}$, then we will also have a “principal bundle isomorphism” from $O(TM, E)$ into $O(TM)|f(M)$. Instead of looking for f directly, we will look for this principal bundle isomorphism. Since the graph of this principal bundle isomorphism is a subset of $O(TM, E) \times O(TN)$, we will look for the graph as an integral submanifold of a certain distribution on $O(TM, E) \times O(TN)$; if $(u, v) \in O(TM, E)$ and $w \in O(TN)$, then the integral manifold through $((u, v), w)$ will turn out to be the graph of the principal bundle isomorphism determined by (f_*, \tilde{f}) , where $f: M \rightarrow N$ is an isometry with $f_*(u_i) = w_i$, and $\tilde{f}(v_r) = w_r$. Thus, although our set-up is now rather complicated, it is very natural to look for maps from $O(TM, E)$ to $O(TN)$, because they involve precisely the right amount of leeway which we expect in the choice of the isometric imbedding $f: M \rightarrow N$.

20. THEOREM. The results of Theorem 19 hold when \mathbb{R}^m is replaced by a complete connected Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 , and the following modifications are made:

- (1) The map A in the conclusion of part (1) is replaced by an isometry $A: N \rightarrow N$.

(2) Gauss' Equation in the hypothesis of part (2) is stated as:

$$\begin{aligned} K_0[\langle\langle X, W \rangle\rangle \langle\langle Y, Z \rangle\rangle - \langle\langle X, Z \rangle\rangle \langle\langle Y, W \rangle\rangle] \\ = \langle\langle R(X, Y)Z, W \rangle\rangle + \{\sigma(X, Z), \sigma(Y, W)\} - \{\sigma(Y, Z), \sigma(X, W)\}. \end{aligned}$$

PROOF. Again we will begin by considering the second part of the theorem. So we are given $(M, \langle\langle \cdot, \cdot \rangle\rangle)$ and the bundle $\varpi: E \rightarrow M$, with metric $\{ \cdot, \cdot \}$, a connection δ compatible with $\{ \cdot, \cdot \}$, and a symmetric section σ of the bundle $\text{Hom}(TM \times TM, E)$. Then δ gives us a connection form $\bar{\Psi}$ on $O(E)$ which is $\mathfrak{o}(m-n)$ -valued; we will denote its components by $\bar{\Psi}_s^r$ for $r, s = n+1, \dots, m$. Similarly, $(\bar{\Psi}_s^r)$ will be the curvature form on $O(E)$.

Now we have obvious maps

$$\begin{array}{ccc} & O(TM, E) & \\ \lambda_1 \swarrow & & \searrow \lambda_2 \\ O(TM) & & O(E). \end{array}$$

For convenience we will denote

$$\begin{array}{ll} \lambda_1^*(\theta^i), \lambda_1^*(\omega_j^i), \lambda_1^*(\Omega_j^i) & \text{simply by } \theta^i, \omega_j^i, \Omega_j^i \\ \lambda_2^*(\bar{\Psi}_s^r), \lambda_2^*(\bar{\Psi}_s^r) & \text{simply by } \bar{\Psi}_s^r, \bar{\Psi}_s^r. \end{array}$$

We define functions $s_{ij}^r: O(TM, E) \rightarrow \mathbb{R}$ as follows. An element of $O(TM, E)$ is a pair (u, v) , where u and v are orthonormal frames, of TM and E , respectively, at the same point p . Then $\sigma(u_i, u_j)$ can be written uniquely as

$$\sigma(u_i, u_j) = \sum_r s_{ij}^r ((u, v)) \cdot v_r.$$

We now define forms $\bar{\Psi}_i^r$ directly on $O(TM, E)$ by

$$\bar{\Psi}_i^r = \sum_j s_{ij}^r \theta^j \quad \left(= \sum_j s_{ij}^r \lambda_1^*(\theta^j) \right).$$

The symmetry of σ implies that $s_{ij}^r = s_{ji}^r$, and thus that

$$(1) \quad \sum_i \bar{\Psi}_i^r \wedge \theta^i = 0.$$

Now we claim that on the bundle $O(TM, E)$ we have

$$(2) \quad K_0[\theta^i \wedge \theta^j] = \Omega_j^i - \sum_r \bar{\Psi}_i^r \wedge \bar{\Psi}_j^r$$

$$(3) \quad d\bar{\Psi}_j^r = - \sum_i \bar{\Psi}_i^r \wedge \omega_j^i - \sum_s \bar{\Psi}_s^r \wedge \bar{\Psi}_j^s$$

$$(4) \quad d\bar{\Psi}_s^r = \sum_i \bar{\Psi}_i^r \wedge \bar{\Psi}_i^s - \sum_w \bar{\Psi}_w^r \wedge \bar{\Psi}_s^w.$$

The proof is in two steps. First, consider a local section $(X_1, \dots, X_n, \nu_{n+1}, \dots, \nu_m) = \xi: U \rightarrow O(TM, E)$ on $U \subset M$. Denote $\xi^*(\theta^i)$ by θ^i , etc., and $\xi^*(\bar{\Psi}_i^r)$ by $\bar{\Psi}_i^r$. Then the θ^i , ω_j^i , and Ω_j^i are the forms for the moving frame X_1, \dots, X_n . When we apply ξ^* to equations (2)–(4), we obtain equations on M , of which the first, for example, reads

$$(2') \quad K_0[\theta^i \wedge \theta^j] = \Omega_j^i - \sum_r \bar{\Psi}_i^r \wedge \bar{\Psi}_j^r.$$

When we take into account the fact that

$$\bar{\Psi}_i^r = \xi^* \bar{\Psi}_i^r = \sum_k (s_{ik}^r \circ \xi) \cdot \xi^* \theta^k = \sum_k \langle \sigma(X_i, X_k), \nu_r \rangle \theta^k,$$

we find, by a straightforward calculation, that equation (2') is equivalent to Gauss' equation. As in the proof of Theorem 18, we can even avoid the calculation by realizing that it will be essentially the same as the calculation which shows that equation (A') is equivalent to Gauss' equation. In a similar manner, we see that true equations result from applying ξ^* to (3) and (4). This means that equations (2)–(4) hold when applied to tangent vectors which are not vertical. So we just have to prove that (2)–(4) hold when applied to a pair of vectors of which at least one is vertical.

The 1-forms θ^i on $O(TM)$ are zero on any vertical vector, while the 2-forms Ω_j^i are zero on any pair of vectors of which at least one is vertical. Since the vertical vectors of $O(TM, E)$ are precisely the vectors Y for which $\lambda_{1*}(Y)$ is vertical in $O(TM)$ and $\lambda_{2*}(Y)$ is vertical in $O(E)$, we see that the forms θ^i and Ω_j^i on $O(TM, E)$ have the same property as the forms θ^i and Ω_j^i on $O(TM)$. Analogous statements hold for the forms Ψ_s^r on $O(TM, E)$. Moreover, the forms $\bar{\Psi}_i^r$ are clearly 0 on vertical vectors of $O(TM, E)$. It is thus clear that (2) holds when applied to a pair of vectors one of which is vertical. To treat equation (4), we note that the structural equation for $O(E)$ gives

$$d\bar{\Psi}_s^r = - \sum_w \bar{\Psi}_w^r \wedge \bar{\Psi}_s^w + \bar{\Psi}_s^r.$$

Since $\bar{\Psi}_s^r$ is zero on a pair of vectors one of which is vertical, while Ψ_i^r and Ψ_i^s are zero on vertical vectors, this clearly gives the result for equation (4). Thus we are left with equation (3). If *both* our vectors Y_1, Y_2 are vertical, then the right side of (3) is zero; on the other hand, if \tilde{Y}_1, \tilde{Y}_2 are vertical vector fields extending Y_1, Y_2 , then the left side is

$$\begin{aligned} d\bar{\Psi}_j^r(Y_1, Y_2) &= Y_1(\bar{\Psi}_j^r(\tilde{Y}_2)) - Y_2(\bar{\Psi}_j^r(\tilde{Y}_1)) - \bar{\Psi}_j^r([\tilde{Y}_1, \tilde{Y}_2]) \\ &= 0, \end{aligned}$$

since $\bar{\Psi}_j^r$ is zero on vertical vectors. This leaves us with equation (3) in the case where just one vector is vertical.

Every vertical vector at a frame (u, v) at p is $\sigma(\Gamma)(u, v)$ for some Γ in the Lie algebra \mathfrak{g} of the group G of $O(TM, E)$. We want to show that (3) holds when applied to

$$Y_1 = \sigma(\Gamma)(u, v), \quad Y_2 = \xi_*(X_p),$$

for some $X_p \in M_p$ and some local section $(X_1, \dots, X_n, v_{n+1}, \dots, v_m) = \xi: U \rightarrow O(TM, E)$. We extend Y_2 to a vector field \tilde{Y}_2 as follows. First extend X_p to a vector field X on M . Then $\xi_*(X)$ is a vector field defined at just one point in each fibre. We extend $\xi_*(X)$ to \tilde{Y}_2 by making it invariant under R_{a*} for all $a \in G$. This means, in particular, that for the Lie derivative we have

$$0 = L_{\sigma(\Gamma)}\tilde{Y}_2 = [\sigma(\Gamma), \tilde{Y}_2].$$

Equation (3), applied to Y_1, Y_2 thus becomes

$$(3') \quad Y_1(\bar{\Psi}_j^r(\tilde{Y}_2)) = \sum_i \bar{\Psi}_i^r(Y_2)\omega_j^i(Y_1) - \sum_s \bar{\Psi}_s^r(Y_1)\bar{\Psi}_j^s(Y_2).$$

To prove equation (3'), we need information about $R_a^*\bar{\Psi}_j^r$, for $a \in G$. We write a as

$$a = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \quad A \in O(n), \quad B \in O(m-n).$$

Let us note first that the definition of s_{ij}^r gives

$$\sigma((u \cdot A)_i, (u \cdot A)_j) = \sum_r s_{ij}^r((u, v) \cdot a) \cdot (v \cdot B)_r.$$

From this we easily find that

$$s_{ij}^r((u, v) \cdot a) = \sum_{k,l=1}^n \sum_{\rho=n+1}^m A_i^k A_j^l s_{kl}^\rho(u, v) (B^{-1})_\rho^r.$$

We can now compute

$$R_a^* \bar{\Psi}_i^r = \sum_j (s_{ij}^r \circ R_a) \cdot R_a^* \theta^i,$$

using the equation (Proposition II.8-12)

$$R_A^* \theta^i = \sum_k (A^{-1})_k^i \theta^k;$$

we find that the $(m - n) \times n$ matrix $[\bar{\Psi}] = (\bar{\Psi}_i^r)$ satisfies

$$(*) \quad R_a^* [\bar{\Psi}] = B^{-1} [\bar{\Psi}] A.$$

Now suppose we write $\Gamma \in \mathfrak{g}$ as

$$\Gamma = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} \quad \text{for } \Gamma_1 \in \mathfrak{o}(n), \quad \Gamma_2 \in \mathfrak{o}(m - n).$$

An integral curve of $Y_1 = \sigma(\Gamma)(u, v)$ is given by

$$t \mapsto (u, v) \cdot \exp t\Gamma = R_{\exp t\Gamma}(u, v).$$

So for the left side of (3') we find that

$$\begin{aligned} Y_1(\bar{\Psi}_j^r(\tilde{Y}_2)) &= \lim_{h \rightarrow 0} \frac{1}{h} [\bar{\Psi}_j^r(\tilde{Y}_2(R_{\exp h\Gamma}(u, v))) - \bar{\Psi}_j^r(Y_2)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [\bar{\Psi}_j^r(R_{\exp h\Gamma} Y_2) - \bar{\Psi}_j^r(Y_2)] \\ &\quad \text{(from the way } \tilde{Y}_2 \text{ was defined)} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [((R_{\exp h\Gamma})^* \bar{\Psi}_j^r)(Y_2) - \bar{\Psi}_j^r(Y_2)] \\ &= ({}_j^r) \text{ component of} \\ &\quad \lim_{h \rightarrow 0} \frac{1}{h} [(\{\exp h\Gamma_2\}^{-1} \cdot [\bar{\Psi}] \cdot \exp h\Gamma_1)(Y_2) - [\bar{\Psi}](Y_2)] \\ &\quad \text{by equation } (*) \\ &= ({}_j^r) \text{ component of } \{-\Gamma_2 \cdot [\bar{\Psi}(Y_2)] + [\bar{\Psi}(Y_2)]\Gamma_1\} \\ &= -\sum_s (\Gamma_2)_s^r \bar{\Psi}_j^s(Y_2) + \sum_i \bar{\Psi}_i^r(Y_2) \cdot (\Gamma_1)_j^i \\ &= -\sum_s \bar{\Psi}_s^r(\sigma(\Gamma)) \bar{\Psi}_j^s(Y_2) + \sum_i \bar{\Psi}_i^r(Y_2) \omega_j^i(\sigma(\Gamma_1)) \\ &= -\sum_s \bar{\Psi}_s^r(Y_1) \bar{\Psi}_j^s(Y_2) + \sum_i \bar{\Psi}_i^r(Y_2) \omega_j^i(Y_1), \end{aligned}$$

which is precisely the right side of (3').

After all this work, we are ready to construct a distribution on the product $O(TM, E) \times O(TN)$. We introduce the two projections

$$\begin{array}{ccc} O(TM, E) \times O(TN) & \xrightarrow{\pi_2} & O(TN) \\ \downarrow \pi_1 & & \\ O(TM, E) & & \end{array}$$

Define $\Delta_{((u,v),w)}$ to be the set of all tangent vectors at $((u,v),w)$ on which the following 1-forms all vanish:

- (a) $\pi_2^* \phi^i - \pi_1^* \theta^i$
- (b) $\pi_2^* \phi^r$
- (c) $\pi_2^* \psi_j^i - \pi_1^* \omega_j^i$
- (d) $\pi_2^* \psi_s^r - \pi_1^* \bar{\psi}_s^r$
- (e) $\pi_2^* \bar{\psi}_j^r - \pi_1^* \bar{\psi}_j^r.$

Since the forms $\phi^\alpha, \psi_\beta^\alpha$ ($\alpha < \beta$) are a basis for the dual space of the tangent space $O(TN)_w$, it is clear that each $\Delta_{((u,v),w)}$ has dimension

$$\dim O(TM, E) \times O(TN) - \dim O(TN) = \dim O(TM, E),$$

and that $\pi_{1*}: \Delta_{((u,v),w)} \rightarrow O(TM, E)_{(u,v)}$ is an isomorphism. We thus obtain a distribution Δ .

For any $a \in G$, we have the map $R_a: O(TM, E) \rightarrow O(TM, E)$. Since $a \in G \subset O(m) = \text{group of the bundle } O(TN)$, we also have maps, which we will denote by the same letter, $R_a: O(TN) \rightarrow O(TN)$. These maps then give us maps

$$R_a: O(TM, E) \times O(TN) \rightarrow O(TM, E) \times O(TN).$$

We claim that $R_{a*} \Delta = \Delta$. To prove this, it suffices to show that $R_a^* \eta$ is a linear combination of the forms (a)–(e) whenever η is any of the forms in (a)–(e). If

$$a = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix},$$

then Proposition II.8-12 shows that the \mathbb{R}^m -valued form $\pi_2^* \phi$ satisfies

$$R_a^* \pi_2^* \phi = \pi_2^* R_a^* \phi = \pi_2^* a^{-1} \phi = \pi_2^* \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix} \phi,$$

while

$$R_a^* \pi_1^* \theta = \pi_1^* R_a^* \theta = \pi_1^* A^{-1} \theta.$$

From this we see that $R_a^* \eta$ has the required property when η is one of the forms in (a) or (b). We also have, by Proposition II.8-11,

$$R_a^* \pi_2^* \psi = \pi_2^* R_a^* \psi = \pi_2^* a^{-1} \psi a = \pi_2^* \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix} \psi \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix};$$

for the same reason we have

$$\begin{aligned} R_a^* \pi_1^* \omega &= \pi_1^* A^{-1} \omega A, \\ R_a^* \pi_1^* \bar{\psi} &= \pi_1^* B^{-1} \bar{\psi} B, \quad \bar{\psi} = (\bar{\psi}_s^r), \end{aligned}$$

while (*) gives

$$R_a^* \pi_1^* [\bar{\psi}] = \pi_1^* B^{-1} [\bar{\psi}] A \quad [\bar{\psi}] = (\bar{\psi}_i^r).$$

From these equations we see that $R_a^* \eta$ has the required property when η is one of the forms in (c)–(e).

Now we claim that our distribution Δ is *integrable*. According to Proposition I.7-14, we just have to show that the differentials of all the forms in (a)–(e) are in the ideal \mathcal{I} generated by these forms. Now we have, for example,

$$\begin{aligned} d(\pi_2^* \phi^i - \pi_1^* \theta^i) &= \pi_2^* \left(- \sum_j \psi_j^i \wedge \phi^j \right) - \pi_1^* \left(- \sum_j \omega_j^i \wedge \theta^j \right) \\ &\quad + \pi_2^* \left(- \sum_r \psi_r^i \wedge \phi^r \right). \end{aligned}$$

The last term is in \mathcal{I} , since the forms (b) are. The first two terms can be written

$$- \sum_j \pi_2^* \psi_j^i \wedge (\pi_2^* \phi^j - \pi_1^* \theta^j) - \sum_j (\pi_2^* \psi_j^i - \pi_1^* \omega_j^i) \wedge \pi_1^* \theta^j,$$

which is in \mathcal{I} . For the exterior differential

$$d(\pi_2^* \phi^r) = \pi_2^* \left(- \sum_i \psi_i^r \wedge \phi^i \right) + \pi_2^* \left(- \sum_s \psi_s^r \wedge \phi^s \right).$$

we note that the second term is in \mathcal{I} , while the first can be written

$$\begin{aligned} - \sum_i (\pi_2^* \psi_i^r - \pi_1^* \bar{\psi}_i^r) \wedge \pi_2^* \phi^i - \sum_i \pi_1^* \bar{\psi}_i^r \wedge (\pi_2^* \phi^i - \pi_1^* \theta^i) \\ - \sum_i \pi_1^* (\bar{\psi}_i^r \wedge \theta^i); \end{aligned}$$

the first two terms are in \mathcal{I} , while the third is zero by equation (1). We will briefly outline the check for (c), and leave the others for the reader. We have

$$\begin{aligned} d(\pi_2^* \psi_j^i - \pi_1^* \omega_j^i) &= \pi_2^* \left(- \sum_k \psi_k^i \wedge \psi_j^k \right) - \pi_1^* \left(- \sum_k \omega_k^i \wedge \omega_j^k \right) \\ &\quad + \pi_2^* \left(- \sum_r \psi_r^i \wedge \psi_j^r \right) + \pi_2^*(\Psi_j^i) - \pi_1^*(\Omega_j^i). \end{aligned}$$

Since N has constant curvature K_0 , we have $\pi_2^*(\Psi_j^i) = K_0 \pi_2^*(\phi^i \wedge \phi^j)$, while equation (2) tells us how to get rid of $\pi_1^*(\Omega_j^i)$. We obtain, in particular, the term

$$\begin{aligned} K_0[\pi_2^* \phi^i \wedge \pi_2^* \phi^j - \pi_1^* \theta^i \wedge \pi_1^* \theta^j] \\ = K_0[(\pi_2^* \phi^i - \pi_1^* \theta^i) \wedge \pi_2^* \phi^j \\ + \pi_1^* \theta^i \wedge (\pi_2^* \phi^j - \pi_1^* \theta^j)], \end{aligned}$$

which is in \mathcal{I} . The other terms are easily paired off and treated as above.

Now consider an integral manifold Γ of the distribution Δ . Since the map $\pi_{1*}: \Delta_{((u,v),w)} \rightarrow O(TM, E)_{(u,v)}$ is always an isomorphism, the map $\pi_1: \Gamma \rightarrow O(TM, E)$ is a diffeomorphism in a neighborhood of any point. Replacing M by a sufficiently small open subset of M if necessary, we may assume that $\pi_1: \Gamma \rightarrow O(TM, E)$ is a diffeomorphism. Then Γ is the graph of a function $g: O(TM, E) \rightarrow O(TN)$, given explicitly by

$$g = \pi_2 \circ (\pi_1|_{\Gamma})^{-1}.$$

Because $R_{a*}\Delta = \Delta$ for all $a \in G$, it is easy to see that g takes fibres of $O(TM, E)$ to fibres of $O(TN)$, so that there is a diffeomorphism $f: M \rightarrow N$ for which the following diagram commutes.

$$\begin{array}{ccc} O(TM, E) & \xrightarrow{g} & O(TN) \\ \pi \downarrow & & \downarrow \pi_N \\ M & \xrightarrow{f} & N \end{array}$$

Now suppose we have a tangent vector $X_p \in M_p$, a frame $(u, v) \in \pi^{-1}(p)$, and a tangent vector $Y \in O(TM, E)_{(u,v)}$ with $\pi_* Y = X_p$. Then, by definition of θ^i , we have

$$\theta^i(Y) = i^{\text{th}} \text{ component of } X_p \text{ with respect to the frame } u.$$

Now

$$f_*X_p = f_*\pi_*Y = \pi_{N*}g_*Y,$$

so we likewise have

$$\begin{aligned} & i^{\text{th}} \text{ component of } f_*X_p \text{ with respect to the frame } g((u, v)) \\ &= \phi^i(g_*Y) \\ &= \phi^i(\pi_{2*}(\pi_1|\Gamma)^{-1}_*(Y)) \\ &= \pi_{2*}\phi^i((\pi_1|\Gamma)^{-1}_*(Y)) \\ &= \pi_{1*}\theta^i((\pi_1|\Gamma)^{-1}_*(Y)) \quad \text{since the forms (a) are zero on } \Gamma \\ &= \theta^i(Y). \end{aligned}$$

Similarly, since the forms (b) vanish on Γ , we find that

$$r^{\text{th}} \text{ component of } f_*X_p \text{ with respect to the frame } g((u, v)) = 0.$$

This shows us that g is of the form

$$g((u, v)) = (f_*(u), \tilde{f}(v)),$$

for some bundle isomorphism $\tilde{f}: E \rightarrow \{\text{normal bundle of } f(M) \text{ in } N\}$ covering f . The map f is an isometry, since f_* takes orthonormal frames to orthonormal frames, and the map \tilde{f} is inner product preserving, for the same reason.

The proof that \tilde{f} makes σ correspond to s and δ correspond to D is similar to the above arguments, using the fact that the forms (c)–(e) vanish on Γ .

We have thus proved the existence part of the theorem locally. Simple-connectivity is then used to prove the global result, in the standard way. The uniqueness part of the theorem is handled just like the uniqueness part of Theorem 18. ♦

For ease of reference, we want to have an explicit statement of Theorem 20 in the case of hypersurfaces. We will assume that the ambient space N is orientable. In this case we claim that a diffeomorphism $\phi: M \rightarrow \bar{M}$ between immersed hypersurfaces is always covered by an inner product preserving bundle isomorphism $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$. To construct $\tilde{\phi}$ we first choose a particular orientation for N . Then for $p \in M$ we choose an ordered basis $X_1, \dots, X_n \in M_p$ and a unit normal $v_p \in M_p^\perp$ such that (X_1, \dots, X_n, v_p) is positively oriented in N_p . Then there is a unique unit normal $\bar{v}_{\phi(p)} = M_{\phi(p)}^\perp$ such that $(\phi_*X_1, \dots, \phi_*X_n, \bar{v}_{\phi(p)})$ is positively oriented in $N_{\phi(p)}$. We let $\tilde{\phi}: M_p^\perp \rightarrow \bar{M}_{\phi(p)}^\perp$ be the linear map taking v_p to $\bar{v}_{\phi(p)}$; it is clear that this map is well-defined. If $-1: \text{Nor } M \rightarrow \text{Nor } M$ is the bundle equivalence taking $X \in M_p^\perp$

to $-X \in M_p^\perp$, then $\tilde{\bar{\phi}} = \tilde{\phi} \circ -1$: $\text{Nor } M \rightarrow \text{Nor } \bar{M}$ is another inner product preserving bundle isomorphism covering ϕ , and these are clearly the only such. When N is oriented, the hypersurfaces M, \bar{M} are also oriented, and $\phi: M \rightarrow \bar{M}$ is orientation preserving, things are even simpler, for there are unit normal fields ν and $\bar{\nu}$ on M and \bar{M} , determined by their orientations (and the orientation of N), and the obvious $\tilde{\phi}$ to consider is the one taking ν to $\bar{\nu}$.

21. THEOREM. Let $(N^{n+1}, \langle \cdot, \cdot \rangle)$ be an orientable complete connected Riemannian manifold of constant curvature K_0 .

(1) Let M and \bar{M} be connected hypersurfaces of N , let $\phi: M \rightarrow \bar{M}$ be an isometry, and let $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ be one of the two inner product preserving bundle isomorphisms covering ϕ . Suppose that either

$$\bar{s}(\phi_* X, \phi_* Y) = \tilde{\phi}(s(X, Y))$$

for all tangent vectors X, Y at all points of M , or

$$\bar{s}(\phi_* X, \phi_* Y) = -\tilde{\phi}(s(X, Y))$$

for all X, Y . Then ϕ is the restriction of an isometry $A: N \rightarrow N$ with $A_* = \tilde{\phi}$ or $A_* = -\tilde{\phi}$ on $\text{Nor } M$.

(1') Choose an orientation for N , and let M and \bar{M} be connected oriented hypersurfaces of N , with unit normal fields ν and $\bar{\nu}$ determined by their orientations (and the orientation of N), with corresponding second fundamental forms II and $\bar{\text{II}}$. Suppose that $\phi: M \rightarrow \bar{M}$ is an orientation preserving isometry with $\phi^* \bar{\text{II}} = \text{II}$. Then ϕ is the restriction of an orientation preserving isometry $A: N \rightarrow N$ with $A_* \nu = \bar{\nu}$.

(2) Let $(M, \langle \cdot, \cdot \rangle)$ be a simply-connected n -dimensional Riemannian manifold, with covariant differentiation ∇ and curvature tensor R , and let S be a symmetric tensor on M , covariant of order 2. Suppose that S satisfies

(1) Gauss' Equation:

$$\begin{aligned} K_0[\langle X, W \rangle \cdot \langle Y, Z \rangle - \langle X, Z \rangle \cdot \langle Y, W \rangle] \\ = \langle R(X, Y)Z, W \rangle + S(X, Z) \cdot S(Y, W) - S(Y, Z) \cdot S(X, W) \end{aligned}$$

(2) The Codazzi-Mainardi Equations:

$$(\nabla_X S)(Y, Z) = (\nabla_Y S)(X, Z).$$

Then there is an isometric immersion $f: M \rightarrow N$ such that $S = f^* \text{II}$, where II is the second fundamental form on $f(M)$ for some unit normal field ν .

One concluding remark is in order. In Chapter 2 we showed that the Gauss and Codazzi-Mainardi equations for a surface in \mathbb{R}^3 are equivalent to the equations of structure of $O(3)$, and that the Fundamental Theorem of Surface Theory reduces to Theorems I.10-17 and I.10-18 about Lie groups. It is not hard (Problem 15) to show, similarly, that the Gauss, Codazzi-Mainardi, and Ricci equations for a submanifold of \mathbb{R}^m are equivalent to the equations of structure of $O(m)$, and that Theorem 19 reduces to theorems about Lie groups. The Lie group $O(m)$ makes its appearance here because of the fact that the group of Euclidean motions of \mathbb{R}^m is a semi-direct product $\mathbb{R}^m \times O(m)$. For a general Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 , the group of isometries cannot be factored in this way. That is why our proof of Theorem 20 involved the bundle $O(N)$ of orthonormal frames of N . As a matter of fact, the bundle $O(N)$ is the group of isometries of N (as a set), since an isometry is determined by knowing which orthonormal frame $u \in O(N)$ is the image of some fixed orthonormal frame u_0 . Thus we ought to be able to interpret the Gauss, Codazzi-Mainardi, and Ricci equations for a submanifold of N as the equations of structure of $O(N)$, equipped with the appropriate group structure, and Theorem 20 should reduce to Theorems I.10-17 and I.10-18. However, we forbear to enter any further into such considerations.

D. FIRST CONSEQUENCES

We begin by considering hypersurfaces $M^n \subset \mathbb{R}^{n+1}$. In a neighborhood of any point of M there is a unit normal field $v: M^n \rightarrow S^n \subset \mathbb{R}^{n+1}$, unique up to sign, and hence a single second fundamental form $\Pi: M_p \times M_p \rightarrow \mathbb{R}$ (which is defined only up to sign). We also have the map $dv: M_p \rightarrow M_p$ with

$$\begin{aligned} \Pi(X_p, Y_p) &= \langle s(X_p, Y_p), v(p) \rangle \\ &= \langle \nabla'_{X_p} Y, v(p) \rangle \\ &= -\langle \nabla'_{X_p} v, Y_p \rangle \\ &= \langle -dv(X_p), Y_p \rangle. \end{aligned}$$

Thus $-dv: M_p \rightarrow M_p$ is a symmetric linear transformation. As in Chapter 2, we define the **principal directions** at p to be the unit eigenvectors $X_p \in M_p$ for $-dv: M_p \rightarrow M_p$, and we define the **principal curvatures** to be the corresponding eigenvalues. Equivalently, the principal curvatures are the eigenvalues of the symmetric matrix $(\Pi(X_i, X_j)) = (\langle s(X_i, X_j), v(p) \rangle)$ for X_1, \dots, X_n an orthonormal basis of M_p .

Various kinds of curvatures can be defined in terms of the k_i . Since the ordering of the k_i is arbitrary, we obviously want to consider only combinations

of the k_i which are invariant under all permutations of the indices $1, \dots, n$. It is well-known that any polynomial function of n variables t_1, \dots, t_n which is invariant under all permutations of $1, \dots, n$ can be written as a polynomial in the “elementary symmetric functions” $\sigma_1, \dots, \sigma_n$ defined by

$$\begin{aligned}\sigma_1(t_1, \dots, t_n) &= \sum_{i=1}^n t_i, & \sigma_2(t_1, \dots, t_n) &= \sum_{i < j} t_i t_j, \\ \sigma_3(t_1, \dots, t_n) &= \sum_{i < j < k} t_i t_j t_k & \dots & \\ \sigma_n(t_1, \dots, t_n) &= t_1 t_2 \cdots t_n.\end{aligned}$$

These functions are the coefficients, up to sign, of the various powers of x in the polynomial

$$\begin{aligned}P_{t_1, \dots, t_n}(x) &= (x - t_1)(x - t_2) \cdots (x - t_n) \\ &= x^n - \sigma_1(t_1, \dots, t_n)x^{n-1} + \cdots + (-1)^n \sigma_n(t_1, \dots, t_n).\end{aligned}$$

[Recall also that if $\sigma_i(t_1, \dots, t_n) = \sigma_i(u_1, \dots, u_n)$ for all i , then the polynomials $P_{t_1, \dots, t_n}(x)$ and $P_{u_1, \dots, u_n}(x)$ are equal, and thus the set of their roots, $\{t_1, \dots, t_n\}$ and $\{u_1, \dots, u_n\}$, are also equal, counting multiplicities.] We define the **(elementary symmetric) curvatures** $K_1(p), \dots, K_n(p)$ by

$$\binom{n}{j} K_j(p) = \sigma_j(k_1, \dots, k_n),$$

where the k_i are the principal curvatures at p ; the binomial coefficient $\binom{n}{j}$ is inserted for sentimental reasons. In particular,

$$H(p) = K_1(p) = \frac{k_1 + \cdots + k_n}{n} \quad \text{is called the **mean curvature**,}$$

$$K(p) = K_n(p) = k_1 \cdots k_n \quad \text{is called the **Gaussian curvature**.}$$

Notice that $K_j(p)$ is independent of the choice of v for j even, while $K_j(p)$ is only defined up to sign for j odd.

In the case of surfaces, we found that the Gaussian curvature $K = k_1 \cdot k_2$ is an invariant under isometry. In general, we have

22. PROPOSITION. For hypersurfaces of \mathbb{R}^{n+1} , the set of the $\binom{n}{2}$ numbers $\{k_i k_j : i < j\}$ is invariant under isometry: If $f: M \rightarrow \bar{M}$ is an isometry between two hypersurfaces $M, \bar{M} \subset \mathbb{R}^{n+1}$, and k_1, \dots, k_n are the principal curvatures of M at p , while $\bar{k}_1, \dots, \bar{k}_n$ are the principal curvatures of \bar{M} at $f(p)$, then the sets $\{k_i k_j : i < j\}$ and $\{\bar{k}_i \bar{k}_j : i < j\}$ are equal, counting multiplicities.

PROOF. For $X, Y \in M_p$, let $\tilde{R}(X, Y)$ denote the map $M_p \times M_p \rightarrow \mathbb{R}$ defined by

$$\tilde{R}(X, Y)(Z, W) = \langle R(X, Y)Z, W \rangle.$$

The symmetry properties of R show that the map $\tilde{R}(X, Y)$ is skew-symmetric, so that $\tilde{R}(X, Y) \in \Omega^2(M_p)$. Now the inner product $\langle \cdot, \cdot \rangle_p$ on M_p gives us a map $X \mapsto X^*$ from M_p to M_p^* , defined by $X^*(Y) = \langle X, Y \rangle$. Choose a basis X_1, \dots, X_n for M_p and consider the map from $\Omega^2(M_p)$ to $\Omega^2(M_p)$ given by

$$X_i^* \wedge X_j^* \mapsto \tilde{R}(X_i, X_j);$$

this makes sense since the $X_i^* \wedge X_j^*$ for $i < j$ are a basis for $\Omega^2(M_p)$ and since $\tilde{R}(X_j, X_i) = -\tilde{R}(X_i, X_j)$. We see immediately that under this map

$$(\sum_i a_i X_i)^* \wedge (\sum_j b_j X_j)^* \mapsto \tilde{R}(\sum_i a_i X_i, \sum_j b_j X_j),$$

so we can describe our map, without any choice of basis, as

$$(1) \quad X^* \wedge Y^* \mapsto \tilde{R}(X, Y), \quad \text{from } \Omega^2(M_p) \text{ to } \Omega^2(M_p).$$

Now the vector space $\Omega^2(M_p)$ has dimension $\binom{n}{2}$, so this map has $\binom{n}{2}$ eigenvalues (counting multiplicities). But if X_1, \dots, X_n are principal vectors at p , with corresponding eigenvalues k_1, \dots, k_n , then Gauss' equation tells us that

$$\tilde{R}(X_i, X_j) = -k_i k_j X_i^* \wedge X_j^*.$$

So the set $\{-k_i k_j : i < j\}$ is the set of eigenvalues of the map (1). Since (1) is defined in terms of the curvature tensor R and the metric $\langle \cdot, \cdot \rangle$, this proves that $\{-k_i k_j : i < j\}$ is invariant under isometry. ♦

23. COROLLARY (THEOREMA EGREGIUM). For hypersurfaces in \mathbb{R}^{n+1} , the Gaussian curvature K is invariant under isometry if n is even, and invariant up to sign if n is odd.

PROOF. Observe that

$$K^{n-1} = \left(\prod_{i=1}^n k_i \right)^{n-1} = \prod_{i < j} k_i k_j. \quad \blacklozenge$$

There is another way of reaching this result, which will provide us with an explicit formula for K in terms of R and $\langle \ , \ \rangle$, a formula which will be extremely important in Chapter 13. First we will do a little linear algebra. Let V be a vector space of *even* dimension n , and let $f: V \rightarrow V$ be a linear transformation having matrix $A = (a_{ij})$ with respect to a basis v_1, \dots, v_n . We propose to find $\det f = \det A$ in terms of the determinants of all 2×2 submatrices of A . We will let

$$D(i_1, i_2; j_1, j_2) = a_{i_1 j_1} a_{i_2 j_2} - a_{i_1 j_2} a_{i_2 j_1},$$

so that if $i_1 < i_2$ and $j_1 < j_2$, then $D(i_1, i_2; j_1, j_2)$ is the determinant of the 2×2 submatrix of A obtained by selecting rows i_1 and i_2 , and columns j_1 and j_2 . Recall that $\det f$ can be defined as follows. The linear transformation f gives us a map $f^*: \Omega^k(V) \rightarrow \Omega^k(V)$ defined by

$$f^*(T)(v_1, \dots, v_k) = T(f(v_1), \dots, f(v_k)), \quad \text{all } T \in \Omega^k(V).$$

In particular, we have the map $f^*: \Omega^n(V) \rightarrow \Omega^n(V)$. Since $\Omega^n(V)$ is 1-dimensional, this map must be multiplication by a constant; and this constant is, in fact, just $\det f$. Now our map f^* also satisfies

$$f^*(\phi_1 \wedge \dots \wedge \phi_k) = f^*(\phi_1) \wedge \dots \wedge f^*(\phi_k) \quad \text{all } \phi_i \in \Omega^1(V).$$

In particular, let the ϕ_i be the dual basis to the v_i . Then

$$f(v_i) = \sum_{j=1}^n a_{ji} v_j \implies f^*(\phi_i) = \sum_{j=1}^n a_{ij} \phi_j.$$

So

$$\begin{aligned} f^*(\phi_1 \wedge \dots \wedge \phi_n) &= [f^*(\phi_1) \wedge f^*(\phi_2)] \wedge \dots \\ &= \left(\sum_{j=1}^n a_{1j} \phi_j \wedge \sum_{k=1}^n a_{2k} \phi_k \right) \wedge \dots \\ &= \left(\sum_{j < k} [a_{1j} a_{2k} - a_{1k} a_{2j}] \phi_j \wedge \phi_k \right) \wedge \dots \\ &= \left(\frac{1}{2} \sum_{j, k} D(1, 2; j, k) \phi_j \wedge \phi_k \right) \wedge \dots \end{aligned}$$

From this we see that

$$\det f = \frac{1}{2^{n/2}} \sum_{j_1, \dots, j_n} D(1, 2; j_1, j_2) \cdots D(n-1, n; j_{n-1}, j_n) \varepsilon^{j_1 \cdots j_n},$$

where

$$\varepsilon^{j_1 \dots j_n} = \begin{cases} 1 & j_1, \dots, j_n \text{ is an even permutation of } 1, \dots, n \\ -1 & j_1, \dots, j_n \text{ is an odd permutation of } 1, \dots, n \\ 0 & j_1, \dots, j_n \text{ are not all distinct.} \end{cases}$$

We can clearly also write

$$\det f = \frac{1}{2^{n/2} n!} \sum_{\substack{i_1, \dots, i_n \\ j_1, \dots, j_n}} D(i_1, i_2; j_1, j_2) \cdots D(i_{n-1}, i_n; j_{n-1}, j_n) \varepsilon^{i_1 \dots i_n} \varepsilon^{j_1 \dots j_n}.$$

Now we apply this formula to evaluate

$$K(p) = \det -dv: M_p \rightarrow M_p$$

in terms of a basis X_1, \dots, X_n of M_p . Using Fact 0 from Chapter 2, we have

$$K = \frac{1}{\det(\langle X_i, X_j \rangle)} \cdot \det(\Pi(X_i, X_j)).$$

For the determinants of the 2×2 submatrices of the matrix $(\Pi(X_i, X_j))$ we have, by Gauss' equation,

$$D(i_1, i_2; j_1, j_2) = \langle R(X_{i_2}, X_{i_1})X_{j_1}, X_{j_2} \rangle.$$

So

$$\begin{aligned} K(p) = \frac{1}{2^{n/2} n!} \sum_{\substack{i_1, \dots, i_n \\ j_1, \dots, j_n}} & \langle R(X_{i_2}, X_{i_1})X_{j_1}, X_{j_2} \rangle \\ & \cdots \langle R(X_{i_n}, X_{i_{n-1}})X_{j_{n-1}}, X_{j_n} \rangle \cdot \frac{\varepsilon^{i_1 \dots i_n} \varepsilon^{j_1 \dots j_n}}{\det(\langle X_i, X_j \rangle)}. \end{aligned}$$

If we have a coordinate system x^1, \dots, x^n on M , and let $X_i = \partial/\partial x^i$, then

$$\begin{aligned} \langle R(X_{i_2}, X_{i_1})X_{j_1}, X_{j_2} \rangle &= \left\langle R\left(\frac{\partial}{\partial x^{i_2}}, \frac{\partial}{\partial x^{i_1}}\right) \frac{\partial}{\partial x^{j_1}}, \frac{\partial}{\partial x^{j_2}} \right\rangle \\ &= R_{j_2 j_1 i_2 i_1} \quad (\text{see pg. II.190}) \\ &= R_{i_1 i_2 j_1 j_2}. \end{aligned}$$

So we can write

$$K = \frac{1}{2^{n/2}n!} \sum_{\substack{i_1, \dots, i_n \\ j_1, \dots, j_n}} R_{i_1 i_2 j_1 j_2} \cdots R_{i_{n-1} i_n j_{n-1} j_n} \cdot \frac{\varepsilon^{i_1 \dots i_n}}{\sqrt{\det(g_{ij})}} \cdot \frac{\varepsilon^{j_1 \dots j_n}}{\sqrt{\det(g_{ij})}}.$$

The symbol $\varepsilon^{i_1 \dots i_n} / \sqrt{\det(g_{ij})}$ which appears in this formula has the following natural interpretation. We have a map

$$\underbrace{M_p^* \times \cdots \times M_p^*}_{n \text{ times}} \xrightarrow{\wedge} \Omega^n(M_p)$$

given by

$$(\phi_1, \dots, \phi_n) \mapsto \phi_1 \wedge \cdots \wedge \phi_n.$$

In particular,

$$(dx^{i_1}(p), \dots, dx^{i_n}(p)) \mapsto \varepsilon^{i_1 \dots i_n} \cdot (dx^1(p) \wedge \cdots \wedge dx^n(p)).$$

Now the metric $\langle \cdot, \cdot \rangle_p$ on M_p determines (compare pg. I.311) two elements of norm 1 in the 1-dimensional vector space $\Omega^n(M_p)$, namely

$$\pm \sqrt{\det(g_{ij}(p))} \cdot dx^1(p) \wedge \cdots \wedge dx^n(p).$$

If we choose an orientation for M , then we have a way of choosing between these two elements (choose the $+$ sign if and only if x^1, \dots, x^n is a positively oriented coordinate system), and we therefore have a map $\Omega^n(M_p) \rightarrow \mathbb{R}$ defined by taking this element to 1. The composition

$$\varepsilon: \underbrace{M_p^* \times \cdots \times M_p^*}_{n \text{ times}} \xrightarrow{\wedge} \Omega^n(M_p) \rightarrow \mathbb{R}$$

is then a contravariant vector field of order n , and its components in the x^1, \dots, x^n coordinate system are precisely $\varepsilon^{i_1 \dots i_n} / \sqrt{\det(g_{ij})}$. If we use \mathcal{R} for the tensor

$$\mathcal{R}(X, Y, Z, W) = \langle R(X, Y)Z, W \rangle,$$

we can then write our formula for K as

$$K = \frac{1}{2^{n/2}n!} \cdot \text{contraction of } (\underbrace{\mathcal{R} \otimes \cdots \otimes \mathcal{R}}_{n/2 \text{ times}} \otimes \varepsilon \otimes \varepsilon).$$

A different choice of orientation for M changes ε to $-\varepsilon$, but doesn't change K .

Proposition 22 also shows that K_2 is invariant under isometry, since

$$\binom{n}{2} K_2(p) = \sigma_2(k_1, \dots, k_n) = \sum_{i < j} k_i k_j = \sigma_1(\{k_i k_j : i < j\}).$$

The other elementary symmetric functions of $\{k_i k_j : i < j\}$ are also invariant under isometry, but in general these functions do not have very nice expressions in terms of k_1, \dots, k_n . More interesting is the fact that K_r is invariant under isometry whenever r is even; this follows from the algebraic fact (Problem 16) that the coefficients of *even* powers of λ in the characteristic polynomial $\chi(\lambda)$ of A can always be expressed in terms of the determinants of the 2×2 submatrices of A .

Now let us consider a hypersurface M^n of a general Riemannian manifold $(N^{n+1}, \langle \cdot, \cdot \rangle)$. We still have a unit normal field ν on M , and corresponding second fundamental form Π with

$$\begin{aligned} \Pi(X_p, Y_p) &= \langle s(X_p, Y_p), \nu(p) \rangle \\ &= \langle \nabla'_{X_p} Y, \nu(p) \rangle = -\langle \nabla'_{X_p} \nu, Y_p \rangle \\ &= \langle A_\nu(X_p), Y_p \rangle. \end{aligned}$$

We can define the **principal directions** at p to be the unit eigenvectors for the self-adjoint map $A_\nu : M_p \rightarrow M_p$, and the **principal curvatures** to be the corresponding eigenvalues. Equivalently, the principal curvatures are the eigenvalues of the symmetric matrix $(\Pi(X_i, X_j))$ for X_1, \dots, X_n an orthonormal basis of M_p . We no longer expect the Theorema Egregium to be true in general—even for surfaces, Gauss' equation for the Gaussian curvature involves not only the metric induced on the surface, but also the curvature of N , which varies from point to point. We do obtain a generalization of the Theorema Egregium in the one case where we would expect it:

24. PROPOSITION. Let N^{n+1} be a Riemannian manifold of constant curvature K_0 . Then for hypersurfaces in N , the set $\{k_i k_j : i < j\}$ of products of principal curvatures is invariant under isometry. Consequently, the Gaussian curvature K_n is invariant under isometry if n is even, and invariant up to sign if n is odd.

PROOF. Exactly like the proof of Proposition 22, except that Gauss' equation gives

$$\tilde{R}(X_i, X_j) = -(k_i k_j + K_0) X_i^* \wedge X_j^*,$$

so the set $\{-k_i k_j - K_0 : i < j\}$ is the set of eigenvalues of the map $X^* \wedge Y^* \mapsto \tilde{R}(X, Y)$. ♦

When we consider submanifolds $M \subset N$ of higher codimension, the definitions given previously no longer make sense. However, if we choose any normal vector $\xi \in M_p^\perp$, then we have the map $A_\xi: M_p \rightarrow M_p$, satisfying

$$\langle s(X, Y), \xi \rangle = \langle A_\xi(X), Y \rangle \quad X, Y \in M_p,$$

so we can define the **principal directions** and **principal curvatures** for ξ to be the unit eigenvectors and corresponding eigenvalues for A_ξ ; equivalently, the principal curvatures are the eigenvalues of the symmetric matrix $(\langle s(X_i, X_j), \xi \rangle)$ for X_1, \dots, X_n an orthonormal basis of M_p . We can then define the (**elementary symmetric**) **curvatures** $K_{1;\xi}, \dots, K_{n;\xi}$ by

$$\binom{n}{j} K_{j;\xi} = \sigma_j(k_1, \dots, k_n),$$

where the k_i are the principal curvatures for ξ . We thus have maps

$$M_p^\perp \rightarrow \mathbb{R} \quad \text{given by} \quad \xi \mapsto K_{j;\xi}.$$

The one interesting (and also very important) case arises for the map

$$M_p^\perp \rightarrow \mathbb{R} \quad \text{given by} \quad \xi \mapsto H_\xi = K_{1;\xi}.$$

This map is *linear*, since $A_{\xi+\xi'} = A_\xi + A_{\xi'}$ and since trace is a linear function of matrices. Therefore there is a unique vector $\eta(p) \in M_p^\perp$ such that

$$\langle \eta(p), \xi \rangle = H_\xi = \frac{\text{trace}(\langle s(X_i, X_j), \xi \rangle)}{n} \quad X_1, \dots, X_n \in M_p \text{ orthonormal}$$

for all $\xi \in M_p^\perp$.

This vector $\eta(p)$ is called the **mean curvature normal** at p . In the case of a hypersurface, $\eta(p) = H(p) \cdot \nu(p)$, where ν is the unit normal (changing ν to $-\nu$ changes H to $-H$, so $H \cdot \nu$ is well-defined). In general, if $\nu_{n+1}, \dots, \nu_m \in M_p^\perp$ is an orthonormal basis, then clearly

$$\eta(p) = \sum_{r=n+1}^m H_{\nu_r} \cdot \nu_r.$$

If, moreover, X_1, \dots, X_n are vector fields tangent to M with $X_1(p), \dots, X_n(p)$ an orthonormal basis for M_p , then

$$\begin{aligned} H_\xi &= \frac{1}{n} \text{trace}(\langle s(X_i(p), X_j(p)), \xi \rangle) \\ &= \frac{1}{n} \sum_{i=1}^n \langle s(X_i(p), X_i(p)), \xi \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \langle \nabla'_{X_i(p)} X_i, \xi \rangle. \end{aligned}$$

Consequently,

$$\begin{aligned}\eta(p) &= \sum_{r=n+1}^m H_{v_r} \cdot v_r \\ &= \frac{1}{n} \sum_{r=n+1}^m \sum_{i=1}^n \langle \nabla'_{X_i(p)} X_i, v_r \rangle \cdot v_r,\end{aligned}$$

whence

$$\eta(p) = \frac{1}{n} \perp \left(\sum_{i=1}^n \nabla'_{X_i(p)} X_i \right), \quad X_1(p), \dots, X_n(p) \text{ orthonormal.}$$

The mean curvature H for a hypersurface, and the mean curvature normal field η in general, will play an important role in Chapter 9.

Even though principal directions and curvatures cannot be defined for submanifolds $M \subset N$ of higher codimension, one definition still makes sense. A point $p \in M$ is called an **umbilic** if the principal curvatures for ξ are all equal, for every $\xi \in M_p^\perp$. In other words, each map $A_\xi: M_p \rightarrow M_p$ must be some multiple of the identity, so for each ξ there must be a λ with

$$A_\xi(X) = \lambda X \implies \langle s(X, Y), \xi \rangle = \lambda \cdot \langle X, Y \rangle \quad \text{for all } X, Y \in M_p.$$

It clearly suffices to have

$$A_{v_r}(X) = \lambda_r X$$

for a basis v_{n+1}, \dots, v_m of M_p^\perp .

If p is an umbilic and we choose an orthonormal basis v_{n+1}, \dots, v_m of M_p^\perp and constants $\lambda_{n+1}, \dots, \lambda_m$ with

$$\langle s(X, Y), v_r \rangle = \lambda_r \langle X, Y \rangle \quad \text{for all } X, Y \in M_p,$$

then

$$s(X, Y) = \sum_{r=n+1}^m \langle s(X, Y), v_r \rangle v_r = \langle X, Y \rangle \cdot \left(\sum_{r=n+1}^m \lambda_r v_r \right).$$

This means that for every $\xi \in M_p^\perp$, and every orthonormal basis X_1, \dots, X_n of M_p , we have

$$\langle s(X_i, X_j), \xi \rangle = \delta_{ij} \left\langle \sum_r \lambda_r v_r, \xi \right\rangle,$$

so

$$\begin{aligned} \frac{1}{n} \text{trace}(\langle s(X_i, X_j), \xi \rangle) &= \frac{1}{n} \left\langle \sum_r \lambda_r v_r, \xi \right\rangle \text{trace}(\delta_{ij}) \\ &= \left\langle \sum_r \lambda_r v_r, \xi \right\rangle. \end{aligned}$$

It follows that $\sum_r \lambda_r v_r$ is precisely $\eta(p)$, so we have

$$s(X, Y) = \langle X, Y \rangle \eta(p), \quad \text{at an umbilic } p.$$

When $s: M_p \times M_p \rightarrow M_p^\perp$ is not the zero map we can set

$$\eta(p) = \sum_{r=m+1}^n \lambda_r v_r = \lambda v_*$$

for a unique non-zero $\lambda \in \mathbb{R}$ and unit vector v_* , and for all $X, Y \in M_p$ we have

$$(*) \quad \begin{cases} \langle s(X, Y), v_* \rangle = \lambda \cdot \langle X, Y \rangle \\ \langle s(X, Y), v \rangle = 0 \quad \text{for } \langle v, v_* \rangle = 0. \end{cases}$$

25. LEMMA. Let $(N^m, \langle \cdot, \cdot \rangle)$ be a space of constant curvature K_0 , and for $n \geq 2$ let M^n be a connected immersed submanifold with all points umbilics. Then either $s = 0$ everywhere, so that M is totally geodesic (by Theorems 1-16 and 1-17), or else $\lambda \neq 0$ is constant and M lies in some $(n+1)$ -dimensional totally geodesic submanifold.

PROOF. Suppose that $s(p) \neq 0$, so that $\lambda(p) \neq 0$. In a neighborhood of p we choose an adapted orthonormal moving frame $X_1, \dots, X_n, X_{n+1}, \dots, X_m$ on M with $X_{n+1} = v_*$ at each point. Then for $1 \leq i \leq n$, and X tangent to M we have, by (*),

$$\psi_i^r(X) = \langle \nabla'_X X_i, X_r \rangle = \langle s(X, X_i), X_r \rangle = \begin{cases} \lambda \langle X, X_i \rangle & r = n+1 \\ 0 & r > n+1, \end{cases}$$

which means that on TM we have

$$\begin{aligned} (1) \quad & \psi_i^{n+1} = \lambda \theta^i \\ (2) \quad & \psi_i^r = 0, \quad r > n+1. \end{aligned}$$

From equation (1) and the Codazzi-Mainardi equations we find that on TM we have

$$\begin{aligned} d\lambda \wedge \theta^i + \lambda d\theta^i &= d\psi_i^{n+1} = - \sum_{\alpha} \psi_{\alpha}^{n+1} \wedge \psi_i^{\alpha} \\ &= - \sum_{k=1}^n \lambda \theta^k \wedge \omega_i^k, \end{aligned}$$

while the first structural equation gives

$$d\theta^i = - \sum_{k=1}^n \omega_k^i \wedge \theta^k = - \sum_{k=1}^n \theta^k \wedge \omega_i^k.$$

So we find that

$$d\lambda \wedge \theta^i = 0, \quad 1 \leq i \leq n.$$

Since $n \geq 2$, this implies that $d\lambda = 0$, so that λ is constant in the neighborhood. This argument shows in general that $\{q \in M : \lambda(q) = \lambda(p)\}$ is open. But this set is also closed, and hence all of M . Thus λ is constant.

Now note that equation (2) gives

$$\begin{aligned} 0 &= d\psi_i^r = - \sum_{\alpha} \psi_{\alpha}^r \wedge \psi_i^{\alpha} = -\psi_{n+1}^r \wedge \lambda \theta^i \\ \implies \psi_{n+1}^r &= 0 \quad \text{on } TM, \quad \text{for } r > n+1. \end{aligned}$$

Therefore

$$\begin{aligned} (3) \quad \nabla'_X v_{\star} &= \nabla'_X X_{n+1} = \sum_{j=1}^n \psi_{n+1}^j(X) \cdot X_j = -\lambda \sum_{j=1}^n \langle X, X_j \rangle \cdot X_j \quad \text{by (1)} \\ &= -\lambda X. \end{aligned}$$

We also have

$$(4) \quad \nabla'_X X_i = \sum_{k=1}^n \psi_i^k(X) X_k + \psi_i^{n+1}(X) \cdot v_{\star} \quad i = 1, \dots, n.$$

Let Δ be the $(n+1)$ -dimensional distribution on M with $\Delta(p) = M_p + \mathbb{R} \cdot v_{\star}(p)$. Equations (3) and (4) and Pre-Lemma 7 show that Δ is parallel along every curve c lying in M . So Corollary 11 implies that M lies in an $(n+1)$ -dimensional totally geodesic subspace of N . ♦

For the case $K_0 = 0$, we can immediately characterize the all-umbilic submanifolds:

26. THEOREM. For $n \geq 2$, let $M^n \subset \mathbb{R}^m$ be a connected immersed submanifold of \mathbb{R}^m with all points umbilics. Then either M lies in some n -dimensional plane or else M lies in some n -dimensional sphere in some $(n+1)$ -dimensional plane.

PROOF. We just have to show that if $\lambda \neq 0$ in Lemma 25, then M lies in a sphere of radius $1/\lambda$. We simply repeat the proof from Lemma 1: Let V be the vector field on \mathbb{R}^m defined by

$$V(p) = p_p \in \mathbb{R}^m_p.$$

Then $\nabla'_X V = X$ for all tangent vectors X of \mathbb{R}^m , so we can write equation (3) in Lemma 25 as

$$\nabla'_X (X_{n+1} + \lambda V) = 0.$$

Thus the vector field $X_{n+1} + \lambda V$ is parallel along M . Identifying tangent vectors of \mathbb{R}^m with elements of \mathbb{R}^m , this means that $X_{n+1} + \lambda V$ is a constant vector v_0 on M , so we have

$$X_{n+1}(p) + \lambda \cdot p = v_0 \in \mathbb{R}^m.$$

Thus

$$p = \frac{v_0 - X_{n+1}(p)}{\lambda}$$

for all $p \in M$, which means that M lies in the sphere of radius $1/\lambda$ around the point v_0/λ . ♦

This proof, which depends so strongly on the special properties of \mathbb{R}^m , breaks down completely when we replace \mathbb{R}^m by a complete simply connected manifold of constant curvature $K_0 \neq 0$. Again we have to exploit different descriptions of these manifolds. First we consider the case $K_0 > 0$.

27. THEOREM. Let $S \subset \mathbb{R}^{m+1}$ be an m -sphere. For $n \geq 2$, let M^n be a connected immersed submanifold of S with all points of M umbilics. Then M is part of an n -sphere.

PROOF. We have $M \subset S \subset \mathbb{R}^{m+1}$, with corresponding covariant differentiations $\nabla, \nabla', \nabla''$. Given $X_p, Y_p \in M_p$, extend them to vector fields X, Y in \mathbb{R}^m

which are tangent to M along M , and tangent to S along S . If $\xi \in M_p^\perp \subset S_p$, then

$$\langle \nabla'_{X_p} Y, \xi \rangle = \langle \nabla'_{X_p} Y, \xi \rangle,$$

since $\nabla'_{X_p} Y$ is the component of $\nabla'_{X_p} Y$ tangent to S ; so we have

$$(1) \quad \langle \nabla'_{X_p} Y, \xi \rangle = \lambda \langle X_p, Y_p \rangle \quad \text{for some } \lambda,$$

since p is an umbilic. On the other hand, if $\mathbf{v} \in S_p^\perp \subset \mathbb{R}^{m+1}_p$ is the unit normal, then

$$(2) \quad \langle \nabla'_{X_p} Y, \mathbf{v} \rangle = \frac{1}{r} \langle X, Y \rangle, \quad r = \text{radius of } S,$$

since all points of S are umbilics in \mathbb{R}^{m+1} . Equations (1) and (2) show that all points of M are umbilics when M is considered as a submanifold of \mathbb{R}^{m+1} . Thus the desired result follows immediately from Theorem 26. ♦

Notice that, as predicted by Lemma 25, an n -sphere $\Sigma \subset S$ is either a totally geodesic submanifold of S (when the radius of Σ equals the radius of S), or else is contained in some $(n+1)$ -dimensional totally geodesic submanifold Σ' of S . In the latter case, Σ is a geodesic sphere in Σ' ; thus we have a complete analogy with Theorem 26.

In order to use the same scheme for investigating all-umbilic submanifolds of H^n , we would first have to consider the all-umbilic submanifolds of \mathbb{R}^{n+1} with the Lorentzian metric; these are the planes $P \subset \mathbb{R}^{n+1}$ of various dimensions, and the quadrics

$$Q = \{p \in P : \langle p - p_0, p - p_0 \rangle = c\} \subset P.$$

Then the all-umbilic submanifolds of H^n must be of the form $H^n \cap P$ or $H^n \cap Q$, and we already noted that the latter submanifolds are contained among the former. However, we merely mentioned, but did not prove, the characterization of the sets $H^n \cap P$. So we will use a different method for the case $K_0 < 0$. We have already used the projective model of H^n in the second proof of Lemma 8. Now we will use the conformal model. We appeal to a classical result about conformally equivalent manifolds.

28. PROPOSITION. Let $f: N \rightarrow \bar{N}$ be a conformal equivalence, and let $M \subset N$ be a submanifold of N with an umbilic $p \in M$. Then $f(p)$ is an umbilic of $f(M) \subset \bar{N}$ (but the λ for $f(p)$ need not be the λ for p).

PROOF. Since the result is purely local, we can assume that the underlying spaces of N and \bar{N} are both \mathbb{R}^m , that f is the identity, that $p = f(p) = 0$, and that $M_p = f(M)_{f(p)}$ is the (x^1, \dots, x^n) -plane $\subset \mathbb{R}^m_0$. The metrics for N and \bar{N} have components $g_{\alpha\beta}$ and $\bar{g}_{\alpha\beta}$ satisfying

$$\bar{g}_{\alpha\beta} = e^{2\sigma} g_{\alpha\beta}$$

for some function σ . Then $\bar{g}^{\alpha\beta} = e^{-2\sigma} g^{\alpha\beta}$, and straightforward calculations show that the corresponding Christoffel symbols satisfy the following equations, in which subscripts on σ denote partial derivatives:

$$\begin{aligned} \overline{[\alpha\beta, \gamma]} &= e^{2\sigma} ([\alpha\beta, \gamma] + g_{\alpha\gamma}\sigma_\beta + g_{\beta\gamma}\sigma_\alpha - g_{\alpha\beta}\sigma_\gamma) \\ \bar{\Gamma}_{\alpha\beta}^\gamma &= \Gamma_{\alpha\beta}^\gamma + \delta_\alpha^\gamma\sigma_\beta + \delta_\beta^\gamma\sigma_\alpha - g_{\alpha\beta} \sum_{\mu=1}^m g^{\gamma\mu}\sigma_\mu. \end{aligned}$$

In particular, for $i, j \leq n$ and $r > n$ we have

$$(1) \quad \bar{\Gamma}_{ij}^r = \Gamma_{ij}^r - g_{ij} \cdot \sum_{\mu=1}^m g^{r\mu}\sigma_\mu.$$

The hypothesis that $p = 0$ is an umbilic point for M means that for each $r > n$ there is a constant λ_r with

$$\Gamma_{ij}^r(0) = \lambda_r g_{ij}(0), \quad 1 \leq i, j \leq n.$$

Then equation (1) gives

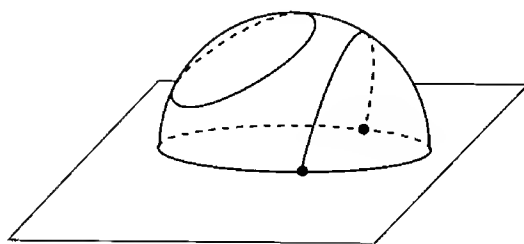
$$\bar{\Gamma}_{ij}^r(0) = \left[\lambda_r - \sum_{\mu=1}^m g^{r\mu}(0)\sigma_\mu(0) \right] \cdot g_{ij}(0),$$

which shows that $f(p) = 0$ is an umbilic for $f(M)$. ♦

29. THEOREM. For $n \geq 2$, let M^n be a connected immersed submanifold of $H^m(K_0)$ with all points of M umbilics. Then either M is totally geodesic, or else M is either a geodesic sphere, a horosphere, or an equidistant hypersurface in some $(n+1)$ -dimensional totally geodesic submanifold of $H^m(K_0)$.

PROOF. Immediate from Lemma 25, Theorem 26, Proposition 28, and our discussion of $(B^m, \langle \cdot, \cdot \rangle)$ in section A. ♦

Proposition 28 could just as well be used to prove Theorem 27. Conversely, if we apply the method used in proving Theorem 27 with the results of Theorem 29, then it is not hard to work backwards and verify the description of geodesic spheres, horospheres, and equidistant hypersurfaces in H^n which was given on page 16. A particular consequence of Theorem 29 is also noteworthy: Any n -sphere contained in H^n , and any n -sphere which intersects \mathbb{R}^{m-1} non-orthogonally, lies in some $(n+1)$ -sphere or $(n+1)$ -plane which intersects \mathbb{R}^{m-1} orthogonally. Presumably one could also hack this result out by elementary



geometry.

For an orthonormal frame X_1, \dots, X_n on an all-umbilic hypersurface $M^{m-1} \subset H^m(K_0)$ ($m \geq 3$) with (constant) λ we have

$$\tilde{R}(X_i, X_j) = -(\lambda^2 + K_0)X_i^* \wedge X_j^* \quad (\text{compare page 70}),$$

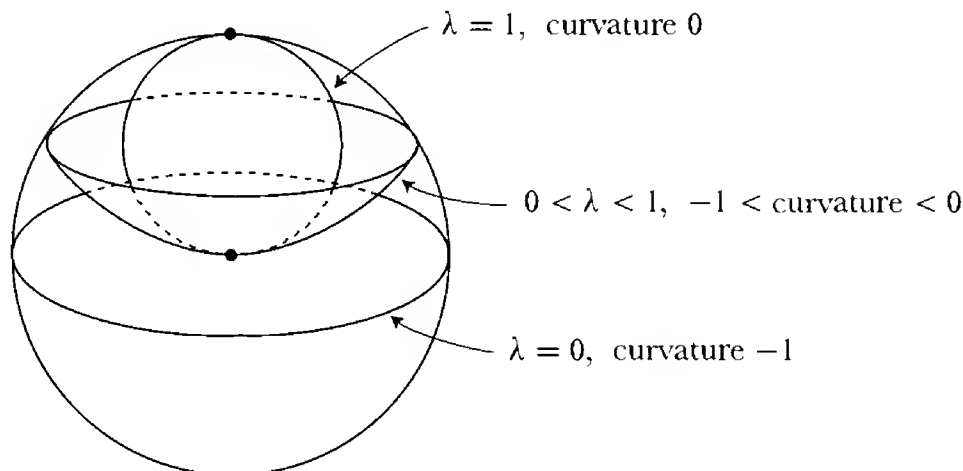
which implies that

$$\begin{aligned} \langle R(X_i, X_j)X_j, X_i \rangle &= \tilde{R}(X_i, X_j)(X_j, X_i) \\ &= \lambda^2 + K_0, \end{aligned}$$

so that M has constant curvature $\lambda^2 + K_0$. Any two all-umbilic hypersurfaces with the same λ are related by an isometry of $H^m(K_0)$, by the first part of Theorem 21. Moreover, there exists a hypersurface with any given $\lambda \geq 0$ (for $\lambda < 0$ we just have the same hypersurface with the opposite choice of unit normal field). In fact, if $(M, \langle \cdot, \cdot \rangle)$ is a simply connected $(m-1)$ -dimensional manifold of constant curvature $\lambda^2 + K_0$, and we define the tensor S on M by $S(X, Y) = \lambda \langle X, Y \rangle$, then M , together with $\langle \cdot, \cdot \rangle$ and S , satisfies Gauss' equation and the Codazzi-Mainardi equations, so by the second part of Theorem 21 there is an isometry of M into $H^m(K_0)$ with second fundamental form Π satisfying $\Pi = \lambda \cdot I$.

It is not hard to determine how the various λ are attached to the various types of all-umbilic hypersurfaces of $H^m(K_0)$. For simplicity, consider $(B^m, \langle \cdot, \cdot \rangle)$, with constant curvature $K_0 = -1$. We know that the horospheres have constant

curvature $0 = \lambda^2 - 1 \implies \lambda = 1$, while the totally geodesic hypersurfaces have constant curvature $-1 = \lambda^2 - 1 \implies \lambda = 0$. We can take a family of all-



umbilic hypersurfaces passing continuously from a totally geodesic hypersurface to a horosphere, with all members of the family distinct up to isometry of B^m . The intermediate hypersurfaces will be equidistant hypersurfaces, and include all such hypersurfaces (up to isometry of B^m). The corresponding λ 's must vary monotonically from 0 to 1. This shows that equidistant hypersurfaces, and only equidistant hypersurfaces, have $0 < \lambda < 1$. So all $\lambda > 1$ must occur for the geodesic spheres. If λ_r is the λ for the geodesic sphere of radius r around 0, then $r \mapsto \lambda_r$ must be a monotonic function of r . Clearly $\lambda_r \rightarrow \infty$ as $r \rightarrow 0$, and $\lambda_r \rightarrow 1$ as $r \rightarrow \infty$.

We have now generalized essentially the material in Chapter 2 which precedes the discussion of the third fundamental form. The facts about higher fundamental forms in general will be left to the Problems. The next generalization on our agenda is then the following.

30. PROPOSITION. If M^n is a compact submanifold immersed in \mathbb{R}^m , then there is a point $p \in M$ and a normal $\xi \in M_p^\perp$ for which the map $A_\xi: M_p \rightarrow M_p$,

$$\langle A_\xi(X), Y \rangle = \langle s(X, Y), \xi \rangle, \quad X, Y \in M_p,$$

is positive definite. $\langle A_\xi(X), X \rangle > 0$ for $X \neq 0$. So if M is a compact hypersurface, with unit normal field ν , then there is a point $p \in M$ for which $-d\nu: M_p \rightarrow M_p$ is either positive or negative definite (depending on the choice of ν). In particular, the Gaussian curvature $K_n(p)$ is non-zero, and in fact $K_n(p) > 0$ for n even.

PROOF. As in the proof of Proposition 2-8, let p be a point of M furthest from 0. Then the line from 0 to p is normal to M at p , and we choose ξ to be the unit vector in M_p pointing in this direction. The rest of the argument is left as an exercise for the reader. ♦

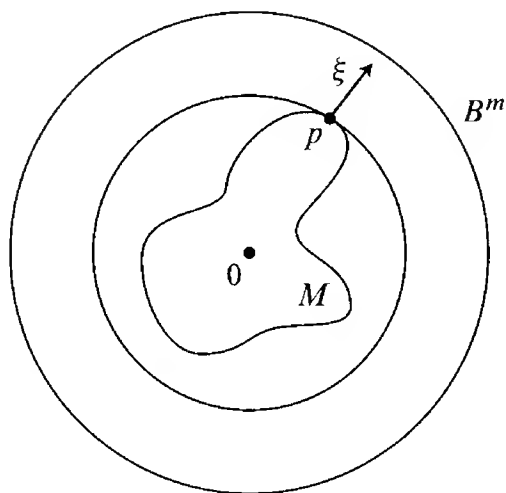
31. COROLLARY. There are no compact submanifolds M^n immersed in \mathbb{R}^m with mean curvature normal $\eta = 0$. In particular, there are no immersed hypersurfaces in \mathbb{R}^m with mean curvature $H = 0$.

PROOF. If ξ is a normal given by Proposition 30, then

$$\langle \eta(p), \xi \rangle = H_\xi = \text{trace } A_\xi,$$

and $\text{trace } A_\xi > 0$ since A_ξ is positive definite. ♦

Suppose we replace \mathbb{R}^m in Proposition 30 by the space $(B^m, \langle \cdot, \cdot \rangle)$ of constant curvature $K_0 < 0$. If $p \in M$ is a point furthest from 0, then M is contained in the geodesic sphere around 0 which passes through p . All principal curvatures



of this sphere are equal to some $\lambda > \sqrt{-K_0}$ (compare pg. III.64). The geodesic from 0 to p is normal to M at p , and if we choose ξ to be the unit normal in M_p^\perp pointing in this direction, then we will have

$$\langle A_\xi(X), X \rangle \geq \lambda > \sqrt{-K_0}.$$

For a hypersphere M , and a correctly chosen unit normal field v , we thus find that all principal curvatures k_1, \dots, k_n are $\geq \lambda > \sqrt{-K_0}$. Hence

$$K_n(p) = \prod_{i=1}^n k_i \geq \lambda^n > (\sqrt{-K_0})^n.$$

In particular, for n even this holds for either choice of v . We also see that there are no compact immersed submanifolds of N with mean curvature normal $\eta = 0$, and hence no compact immersed hypersurfaces of N with mean curvature $H = 0$.

Now let us replace \mathbb{R}^m by a sphere S of radius $1/\sqrt{K_0}$, for $K_0 > 0$, and suppose moreover, that M is contained in an open hemisphere of S , say the hemisphere centered around the point x . By choosing a point $p \in M$ furthest from x , and a unit normal ξ in M_p pointing along the geodesic from x to p , we find that

$$\langle A_\xi(X), X \rangle \geq \lambda$$

for some $\lambda > 0$. But there is obviously no positive lower bound for all λ 's. For hypersurfaces M we find that $K_n(p) \neq 0$, and $K_n(p) > 0$ for n even, but again there is no positive lower bound for K_n . Similarly, we find that Corollary 31 generalizes to compact submanifolds of an open hemisphere. Naturally, our results break down if we replace the hemisphere by the whole sphere, for the equatorial $(m-1)$ -sphere has second fundamental form $s = 0$. You might think that this is the only exception, but there are actually many other possibilities. In fact, we easily compute that for $p, q \geq 1$ with $p + q = m - 1$, the hypersurface

$$M = \left\{ (x_1, \dots, x_{p+1}, y_1, \dots, y_{q+1}) \in \mathbb{R}^{m+1} : \sum x_k^2 = \frac{p}{m-1} \text{ and } \sum y_k^2 = \frac{q}{m-1} \right\} \\ \subset S^m$$

has mean curvature $H = 0$ in the unit sphere S^m , so there is certainly no point $p \in M$ where $dv: M_p \rightarrow M_p$ is definite.

To complete our generalization of the material in Chapter 2, we want to discuss the relationship between positive curvature and convexity of hypersurfaces. For a hypersurface $M^n \subset \mathbb{R}^{n+1}$, the proper analogue of positivity of the Gaussian curvature at p is the condition that all sectional curvatures at p are positive; equivalently, all principal curvatures should have the same sign, or yet again, the map $dv: M_p \rightarrow M_p$ should be (positive or negative) definite. It is easy to see that definiteness of $dv: M_p \rightarrow M_p$ implies that M locally lies on one side of the tangent hyperplane of M at p . If $dv: M_p \rightarrow M_p$ is merely semi-definite (that is, $\langle dv(X), X \rangle \geq 0$ for all X , or $\langle dv(X), X \rangle \leq 0$ for all X), then no conclusion can be drawn. But if $dv: M_p \rightarrow M_p$ is *not* semi-definite, then M locally lies on both sides of its tangent hyperplane at p . Propositions 2-9 and 2-10 clearly generalize to hypersurfaces in \mathbb{R}^{n+1} ; we will not bother to write down all the details, but will henceforth use the word "convex" for a hypersurface in either of its two equivalent meanings.

32. PROPOSITION.

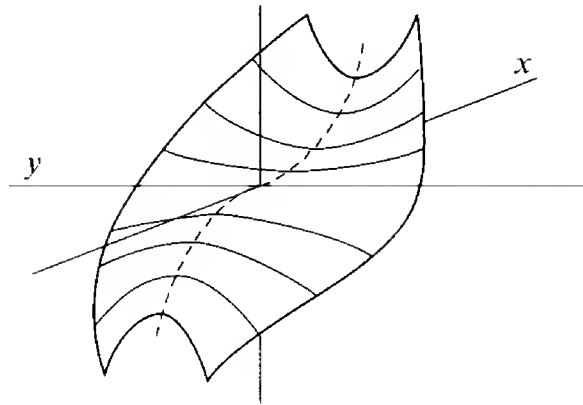
(1) If M is a convex hypersurface in \mathbb{R}^{n+1} , then $dv: M_p \rightarrow M_p$ is semi-definite for all $p \in M$.

(2) Let M be a compact connected n -manifold, and $f: M \rightarrow \mathbb{R}^{n+1}$ an immersion with normal map n such that $dn: M_p \rightarrow M_p$ is definite for all $p \in M$. Then

- (i) The manifold M is orientable, and the normal map $n: M \rightarrow S^n \subset \mathbb{R}^{n+1}$ is a diffeomorphism.
- (ii) The map $f: M \rightarrow \mathbb{R}^{n+1}$ is an imbedding, and $f(M)$ is convex.

PROOF. This generalization of Hadamard's Theorem (2-11) is proved in exactly the same way as the original. ♦

The most significant part of this result is the fact that the immersion f must be an imbedding. In fact, the definiteness of dv implies that M is locally convex, and there are general arguments to show that a locally convex set in \mathbb{R}^m is actually convex, which implies the theorem for an imbedded hypersurface M . On the other hand, we have already mentioned in Chapter 2 that for $n = 2$ Hadamard's Theorem holds even under the weakened assumption that $K(p) \geq 0$ for all $p \in M$. Here the result is not clear even for imbedded $M \subset \mathbb{R}^3$, since the condition $K \geq 0$ does not imply local convexity for arbitrary (non-compact) M . For example, the graph of $(x, y) \mapsto x^3(1 + y^2)$ has



$K \geq 0$ in a neighborhood of $0 \in \mathbb{R}^3$ (by an easy calculation), but is clearly not locally convex. The extension of Hadamard's Theorem for $K \geq 0$ (and $n = 2$) was originally proved by Chern and Lashof [1], using a little Morse theory. Sacksteder [1] then gave a proof for all n under the weakened assumption that $dn: M_p \rightarrow M_p$ is semi-definite for all $p \in M$; in fact, compactness of M can

be replaced by completeness, provided that there is at least one point $p \in M$ where at least one sectional curvature is non-zero (without this last condition, M might be a generalized cylinder). Sacksteder's proof is more "elementary", and, as one might guess, much harder. (For the case where all $dn: M_p \rightarrow M_p$ are definite, but M is merely complete and immersed, there is an earlier proof by Stoker [1].) Do Carmo and Lima [1] gave a simple proof of a result even more general than Sacksteder's when M is compact: If $f: M^n \rightarrow \mathbb{R}^m$ is an immersion with all maps $A_\xi: M_p \rightarrow M_p$ semi-definite (for $\xi \in M_p^\perp$) for all $p \in M$, and all maps A_ξ definite for at least one $p \in M$ [for $m = n + 1$ this latter condition follows from Proposition 30], then $f(M)$ is contained in some $(n + 1)$ -dimensional plane in \mathbb{R}^m , and f is an imbedding of M as a convex set. In Do Carmo and Lima [2], they also give a simple argument which reproves Sacksteder's result for complete M (but which does not recapture all of the additional information obtained in the course of Sacksteder's analysis).

We can also consider convex sets in spaces of constant curvature K_0 . For $H^m(K_0)$, the definition is precisely the same as for \mathbb{R}^m : a set $A \subset H^m(K_0)$ is convex if A contains the segment of the unique geodesic between p and q whenever $p, q \in A$. For $K_0 > 0$, we consider only an open hemisphere of $S^m(K_0)$, so that there is a unique geodesic between any two points, and the same definition can be used. Since geodesic mappings preserve convexity, we see immediately that Proposition 2-10 generalizes when we replace the tangent plane of M at p by the totally geodesic hypersurface $\exp(M_p)$. Again we will use "convex" for hypersurfaces in either of its two equivalent meanings. It also looks as if we should be able to use geodesic mappings to generalize Proposition 32 to hypersurfaces of $H^{n+1}(K_0)$ and $S^{n+1}(K_0)$. The details of this program turn out to be a little sticky, and since the arguments have been covered in a recent paper, Do Carmo and Warner [1], we will merely quote their results:

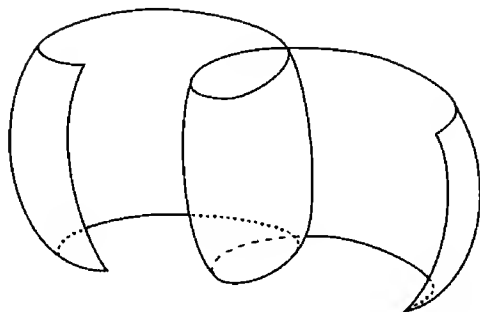
33. THEOREM (DO CARMO-WARNER).

(1) If M is a convex hypersurface in $H^{n+1}(K_0)$ for $K_0 < 0$, or a convex hypersurface in a hemisphere of $S^{n+1}(K_0)$ for $K_0 > 0$, then all sectional curvatures of M are $\geq K_0$. Moreover, if ϕ is a geodesic mapping from $H^{n+1}(K_0)$, or a hemisphere of $S^{n+1}(K_0)$, to \mathbb{R}^{n+1} , then all sectional curvatures of M are $> K_0$ at p if and only if all sectional curvatures of $\phi(M)$ are > 0 at $\phi(p)$.

(2) Let M be a compact connected n -manifold, and $f: M \rightarrow S^{n+1}(K_0)$ an immersion, for $K_0 > 0$, such that all sectional curvatures are $\geq K_0$. Then M is orientable, the immersion f is an imbedding, and either $f(M)$ is totally geodesic, or else $f(M)$ is contained in some open hemisphere and is convex.

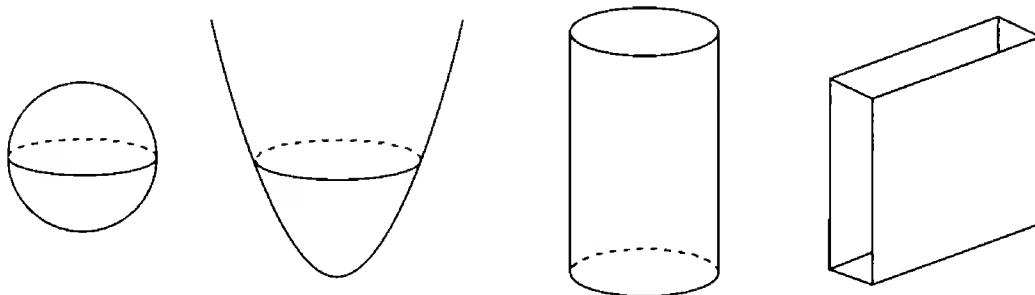
(3) Let M be a compact connected n -manifold, and $f: M \rightarrow H^{n+1}(K_0)$ an immersion, for $K_0 < 0$, such that all sectional curvatures are $\geq K_0$. Then M is orientable, the immersion f is an imbedding, and $f(M)$ is convex.

In part (2) of this result, compactness of M is really equivalent to completeness, by Corollary 8-22. In part (3), compactness does not follow from completeness, and if we try to deal with complete M in $H^{n+1}(K_0)$ we run into the problem that the image $\phi(H^{n+1}(K_0))$ of the geodesic map $\phi: H^{n+1}(K_0) \rightarrow \mathbb{R}^{n+1}$ is an open ball, and hence $\phi \circ f(M)$ need not be complete. As a matter of fact, part (3) is false if M is merely assumed complete. Even if all sectional curvatures of an immersion $f: M \rightarrow H^{n+1}(K_0)$ are $> K_0$, it does not follow that f is an imbedding. To see this, we consider an immersed, but not imbedded,



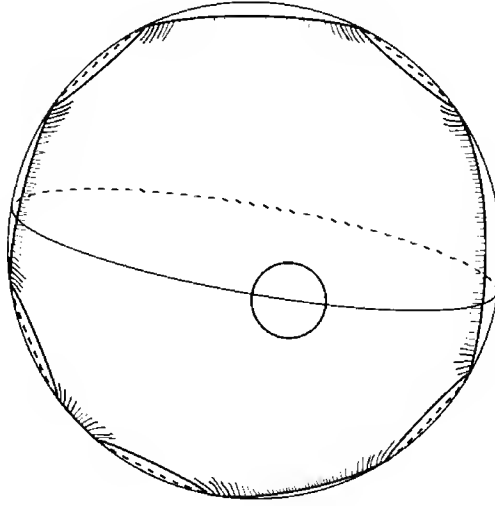
surface in \mathbb{R}^3 with everywhere positive curvature. Such a surface cannot be complete in \mathbb{R}^3 , but its intersection with the projective model of H^3 may very well be complete in H^3 , even though its (extrinsic) curvature is > -1 , by part (1) of Theorem 33. Similarly, if $M \subset \mathbb{R}^3$ is the non-convex surface pictured on page 82, with non-negative curvature near 0, then the intersection of M with the projective model of H^3 can be a complete imbedded surface with extrinsic curvature ≥ -1 everywhere, but it will not be convex in H^3 .

As a concluding remark, we point out that a complete convex hypersurface in \mathbb{R}^{n+1} is of very restricted topological type; it is homeomorphic to S^n (if it



is compact) or to \mathbb{R}^n or $S^1 \times \mathbb{R}^{n-1}$ otherwise. On the other hand, there are

complete convex hypersurfaces of H^{n+1} which are homeomorphic to \mathbb{R}^n with any number of holes, as shown below for the projective model of H^3 .



E. FURTHER RESULTS

This section is devoted to generalizations of certain material in Chapters 3 and 4. The first thing we want to consider are ruled surfaces in \mathbb{R}^m , given by

$$f(s, t) = c(s) + t\delta(s)$$

for two curves c and δ in \mathbb{R}^m . When $m = 3$ we found that the surface is flat precisely when c', δ, δ' are everywhere linearly dependent, by using the Gaussian curvature $k_1 \cdot k_2$, the product of the principal curvatures. For $m > 3$, we have to compute the curvature of the surface f from an intrinsic formula. We can assume that $|\delta| = 1$, and hence $\langle \delta, \delta' \rangle = 0$. Then

$$\left. \begin{array}{l} f_1 = c' + t\delta' \\ f_2 = \delta \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} E = \langle f_1, f_1 \rangle = \langle c', c' \rangle + 2t\langle c', \delta' \rangle + t^2\langle \delta', \delta' \rangle \\ F = \langle f_1, f_2 \rangle = \langle c', \delta \rangle \\ G = 1. \end{array} \right.$$

Most of the terms in the formula on pg. II.129 vanish, and we end up with

$$\begin{aligned} 4(EG - F^2)^2 K &= G \cdot \left(\frac{\partial E}{\partial t} \right)^2 - 2(EG - F^2) \frac{\partial^2 E}{\partial t^2} \\ &= [2\langle c', \delta' \rangle + 2t\langle \delta', \delta' \rangle]^2 \\ &\quad - 2[\langle c', c' \rangle + 2t\langle c', \delta' \rangle + t^2\langle \delta', \delta' \rangle - \langle c', \delta \rangle^2] \cdot 2\langle \delta', \delta' \rangle. \end{aligned}$$

The coefficients of t and t^2 vanish, and we find that

$$K = 0 \iff 0 = \langle c', \delta' \rangle^2 - \langle c', c' \rangle \cdot \langle \delta', \delta' \rangle + \langle c', \delta \rangle^2 \cdot \langle \delta', \delta' \rangle.$$

This condition is automatic when $\delta' = 0$. At points where $\delta' \neq 0$, we can write

$$K = 0 \iff \langle c', c' \rangle = \left\langle c', \frac{\delta'}{|\delta'|} \right\rangle^2 + \langle c', \delta \rangle^2.$$

Since $\delta, \delta'/|\delta'|$ are orthonormal, this happens precisely when c' is a linear combination of δ, δ' . So in all cases,

$$K = 0 \iff c', \delta, \delta' \text{ are linearly dependent.}$$

We can now repeat the analysis on pp. III.236–237 and see that flat ruled surfaces in \mathbb{R}^m are “in general” cylinders, cones, or tangents to a curve.

It should be pointed out that there are plenty of *non*-ruled flat surfaces in \mathbb{R}^m for $m > 3$. For example, the torus

$$S^1 \times S^1 \subset \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4,$$

with the product metric, is flat.

We can also define **ruled surfaces** in an arbitrary Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$. They are the surfaces which can be parameterized as

$$f(s, t) = \exp_{c(s)}(tV(s)),$$

where V is a unit vector field along c .

We want to consider, in particular, the case where N has constant curvature K_0 , and try to describe the ruled surfaces in N which also have constant curvature K_0 . First we consider the case $m = 3$. For a surface $M \subset N^3$ it is important to make a distinction which does not arise in the case of surfaces in \mathbb{R}^3 . The surface M has an induced Riemannian metric, and thus an intrinsic curvature

$$K_{\text{int}}(p) = \langle R(X_p, Y_p)Y_p, X_p \rangle \quad \text{for orthonormal } X_p, Y_p \in M_p.$$

It also has an extrinsic Gaussian curvature $K_{\text{ext}}(p) = k_1 \cdot k_2$, the product of the principal curvatures at p . If N has constant curvature K_0 , then Gauss' equation tells us that

$$(*) \quad K_{\text{int}} = K_{\text{ext}} + K_0.$$

Recall, by the way, that a surface M having constant curvature just means that the function K_{int} on M is constant, while the condition that a higher dimensional manifold have constant curvature is more involved.

The reason for considering the case $m = 3$ first is that in this case the hypothesis that M is ruled is essentially redundant:

34. PROPOSITION. Let N be a 3-dimensional manifold of constant curvature K_0 , and let $M \subset N$ be a surface with constant intrinsic curvature $K_{\text{int}} = K_0$. If $p \in M$ is a point where the second fundamental form $s: M_p \times M_p \rightarrow M_p^\perp$ is not 0, then p has a neighborhood which is a ruled surface.

PROOF. Since M has

$$K_{\text{int}} = K_0 \implies K_{\text{ext}} = 0 \quad \text{by } (*),$$

one principal curvature, k_1 , is always 0. Since s is non-zero at p , the other principal curvature, k_2 , is non-zero in a neighborhood of p . Choose orthonormal vector fields X_1, X_2 on this neighborhood so that each $X_1(q)$ is a principal vector with principal curvature $k_1(q) = 0$, and $X_2(q)$ is a principal vector with principal curvature $k_2(q) \neq 0$. Now the Codazzi-Mainardi equations for N are exactly the same as for \mathbb{R}^3 , so the proof of Proposition 5-4 goes through unchanged, leading to the conclusion that $\nabla'_{X_1} X_1 = 0$, which means that the integral curves of X_1 are geodesics in N . ♦

Naturally, this result does not hold when N has dimension > 3 , so for a general manifold $(N, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 we will now restrict our attention to *ruled* surfaces $M \subset N$. By Synge's inequality (Corollary 1-7) we always have $K_{\text{int}}(p) \leq K_0$. Moreover,

$$K_{\text{int}} = K_0 \text{ along a ruling } \gamma \text{ of } M \iff M_{\gamma(t)} \text{ is parallel along } \gamma.$$

But Lemma 8 shows that

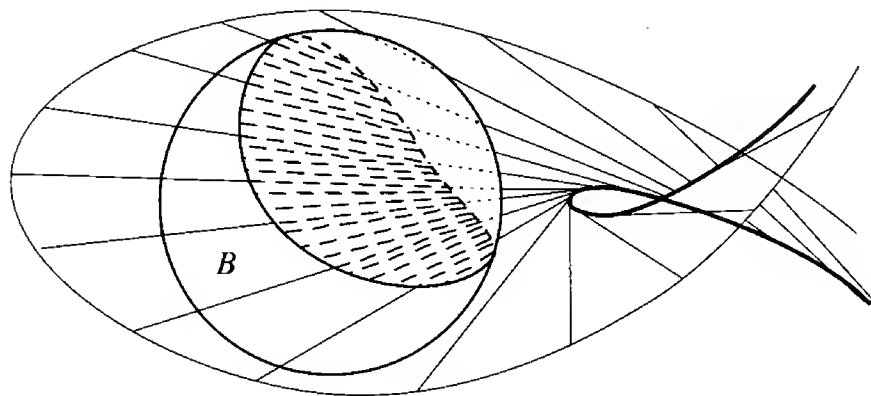
$$M_{\gamma(t)} \text{ is parallel along } \gamma \iff \begin{array}{l} M \text{ is tangent to a} \\ \text{2-dimensional totally geodesic} \\ \text{submanifold of } N \text{ along } \gamma. \end{array}$$

The interesting thing about this last condition is that it does not involve metrics, but only their geodesics. Hence

35. THEOREM. Let N be a manifold of constant curvature K_0 and let $\phi: N \rightarrow \mathbb{R}^m$ be a geodesic mapping. Let $M \subset N$ be a ruled surface. Then M has constant intrinsic curvature $K_{\text{int}} = K_0$ if and only if the ruled surface $\phi(M) \subset \mathbb{R}^m$ is flat.

PROOF. Immediate from the above equivalences. ♦

From Theorem 35 we see that the surfaces $M \subset N$ with $K_{\text{int}} = K_0$ are “in general” ϕ^{-1} of cones, cylinders, and tangent developables. As a local classification, this works equally well for $K_0 < 0$ and $K_0 > 0$. But the situation is quite different when we look for *complete* surfaces with $K_{\text{int}} = K_0$. In the sphere, any pair of geodesics intersect, so there cannot be “cylinders” as in \mathbb{R}^m (this is reflected in the fact that the geodesic mapping from S^m to \mathbb{R}^m is actually defined only on a hemisphere). Once one realizes this, it seems very hard for there to be many such surfaces. In fact, in the next section we will see that in S^3 the only complete surfaces with $K_{\text{int}} = K_0$ are the great 2-spheres. Now consider hyperbolic space H^m . We know that there is a geodesic mapping $\phi: H^m \rightarrow B^m(1)$. Equivalently, there is a metric $\langle \cdot, \cdot \rangle$ on $B^m(1)$ with constant curvature $K_0 < 0$, whose geodesics are just straight lines of \mathbb{R}^m (with a different parameterization). A cone, cylinder, or tangent developable in \mathbb{R}^m then intersects $B^m(1)$ in a surface with $K_{\text{int}} = K_0$ with the metric induced from $\langle \cdot, \cdot \rangle$. The interesting thing is that we can take the vertex of our cone, or the generating curve for the tangent developable to lie *outside* of B . Then



the intersection with B will be a *complete* flat surface, without singularities, of constant intrinsic curvature $K_{\text{int}} = K_0$. Thus there are many such surfaces, of far greater variety than in \mathbb{R}^m . In the next section we will see this in a startling way.

Now consider an oriented surface M in an arbitrary oriented 3-dimensional Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$, and an arclength parameterized curve c in M . We again define the **Darboux frame** of c on M to be the moving frame

$$\mathbf{t}(s) = c'(s), \quad \mathbf{u}(s), \quad \mathbf{v}(s) = \mathbf{t}(s) \times \mathbf{u}(s) = \nu(c(s)),$$

where $\mathbf{u}(s) \in M_{c(s)}$ is a unit vector perpendicular to $\mathbf{t}(s)$ with $(\mathbf{t}(s), \mathbf{u}(s))$ positively oriented in M , and the unit normal field ν is chosen so that the triple

$(\mathbf{t}(s), \mathbf{u}(s), \mathbf{v}(s))$ is positively oriented in N . We still have

$$\begin{aligned}\mathbf{t}' &= \kappa_g \mathbf{u} + \kappa_n \mathbf{v} \\ \mathbf{u}' &= -\kappa_g \mathbf{t} + \tau_g \mathbf{v} \\ \mathbf{v}' &= -\kappa_n \mathbf{t} - \tau_g \mathbf{u}\end{aligned}$$

for certain functions $\kappa_n, \kappa_g, \tau_g$. Everything in Chapter 4 up to and including Proposition 4-5 generalizes almost without any change (asymptotic directions on M are defined just as before, as unit vectors $X \in M_p$ with $\Pi(X_p, X_p) = 0$; they exist only on regions where $K_{\text{ext}}(p) \leq 0$). Moreover, Theorem 4-7 also generalizes, essentially without change. For reference, we merely state this generalization:

36. THEOREM (BELTRAMI-ENNEPER). Let M be a surface in an oriented 3-dimensional Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$. If c is an asymptotic curve in M with $c(0) = p$ and first curvature $\kappa_1(0) \neq 0$, then

$$|\kappa_2(0)| = \sqrt{-K_{\text{ext}}(p)}.$$

Moreover, if $K_{\text{ext}}(p) < 0$ and the two distinct asymptotic curves through p both have non-zero first curvature κ_1 at p , then their second curvatures κ_2 at p are negatives of each other.

The next result generalizes Theorem 4-8.

37. THEOREM. Let N^m be a manifold of constant curvature K_0 , let c be an immersed curve in a hypersurface $M \subset N$, and let S be the ruled surface formed by the geodesics of N which are perpendicular to M along c . Then c is a line of curvature if and only if S has constant intrinsic curvature $K_{\text{int}} = K_0$.

PROOF. Since the result is a local one, we can assume that there is a geodesic mapping $\phi: N \rightarrow \mathbb{R}^m$. The surface S is $\{\exp_{c(s)} t\nu(c(s))\}$, where ν is a unit normal field on M . Hence, identifying tangent vectors of \mathbb{R}^m with elements of \mathbb{R}^m as usual, we have

$$\begin{aligned}\phi(S) &= \{\phi(c(s)) + t\phi_*(\nu(c(s)))\} \\ &= \{\gamma(s) + t\delta(s)\}, \quad \text{say.}\end{aligned}$$

If $\bar{\nabla}$ denotes covariant differentiation in \mathbb{R}^m , then, as in the second proof of Lemma 8, we have

$$\bar{\nabla}_{\phi_* X} \phi_* Y - \phi_*(\nabla'_X Y) = \omega(\phi_* X) \cdot \phi_* Y + \omega(\phi_* Y) \cdot \phi_* X,$$

for some 1-form ω on \mathbb{R}^m . Hence

$$\delta'(s) - \phi_*(\nabla'_{c'(s)}\nu) = \text{a linear combination of } \phi_*(c'(s)) \text{ and } \phi_*(\nu(c(s))).$$

Consequently, we can write

$$(1) \quad \delta'(s) = \phi_*(\nabla'_{c'(s)}\nu) + a\phi_*(c'(s)) + b\phi_*(\nu(c(s))).$$

First suppose that c is a line of curvature, so that $\nabla'_{c'(s)}\nu$ is a multiple of $c'(s)$ for all s . Then equation (1) shows that we can write

$$\begin{aligned} \delta'(s) &= \alpha\phi_*(c'(s)) + b\phi_*(\nu(c(s))) \\ &= \alpha\gamma'(s) + b\delta(s). \end{aligned}$$

So γ', δ, δ' are always linearly independent, and the ruled surface $\phi(S) \subset \mathbb{R}^m$ is flat. Hence S has constant intrinsic curvature $K_{\text{int}} = K_0$ by Theorem 35.

Conversely, if S has constant intrinsic curvature $K_{\text{int}} = K_0$, then $\phi(S)$ is flat, so γ', δ, δ' are always linearly dependent. Then (1) shows that for each s there are numbers A, B, C , not all 0, with

$$(2) \quad Ac'(s) + B\nu(c(s)) + C[\nabla'_{c'(s)}\nu + ac'(s) + b\nu(c(s))] = 0.$$

Clearly $C \neq 0$. Taking the inner product of (2) with $\nu(c(s))$ gives

$$B + Cb = 0,$$

and hence (2) becomes

$$(A + Ca)c'(s) + C\nabla'_{c'(s)}\nu = 0,$$

which shows that c is a line of curvature. ♦

In Eisenhardt [1; pg. 213] this result is verified in a more direct way, by using special coordinates—the Weierstrass coordinates. I like the above proof because it has the strange feature that it uses geodesic mappings even though such mappings preserve neither perpendicularity nor lines of curvature. Once one realizes this, it becomes clear how to generalize the theorem vastly (Problem 19).

F. COMPLETE SURFACES OF CONSTANT CURVATURE

In this section we will classify, so far as possible, the complete constant curvature surfaces in the complete simply-connected 3-dimensional manifolds of constant curvature. First consider a surface M in any 3-dimensional manifold $(N, \langle \cdot, \cdot \rangle)$. By Corollary 4-17, for any point $p \in M$ we can find an imbedding $f: U \rightarrow M$ with $U \subset \mathbb{R}^2$ open and $p \in f(U)$ whose coordinate lines are the lines of curvature, or the asymptotic lines [if $K_{\text{ext}}(p) < 0$]. We want to see what the formulas in the Addendum to Chapter 4 become in these cases. As before, E, F, G are the components of $f^*\langle \cdot, \cdot \rangle$ with respect to the standard coordinate system (s, t) on \mathbb{R}^2 , while l, m, n are the components of $f^*\text{II}$, where II is the second fundamental form of the hypersurface $M \subset N$ for some choice of a unit normal field ν on M . The formula in Problem 4-13 gives the intrinsic curvature K_{int} , so we see that

(A) When the parameter lines of $M^2 \subset N^3$ are orthogonal, we have

$$F = 0$$

$$K_{\text{int}} = -\frac{1}{2\sqrt{EG}} \left[\left(\frac{E_2}{\sqrt{EG}} \right)_2 + \left(\frac{G_1}{\sqrt{EG}} \right)_1 \right].$$

We also know that the Codazzi-Mainardi equations for an ambient manifold of constant curvature are the same as in the Euclidean case, so

(B) When N^3 has constant curvature and the parameter lines of $M^2 \subset N^3$ are lines of curvature, we have

$$l = k_1 E, \quad n = k_2 G, \quad m = 0, \quad F = 0$$

$$l_2 = \frac{E_2}{2} \left(\frac{l}{E} + \frac{n}{G} \right)$$

$$n_1 = \frac{G_1}{2} \left(\frac{l}{E} + \frac{n}{G} \right).$$

(C) When N^3 has constant curvature and the parameter lines of $M^2 \subset N^3$ are asymptotic curves, we have

$$l = n = 0$$

$$m_1 = \frac{\left[\frac{1}{2}(EG - F^2)_1 + FE_2 - EG_1 \right]}{EG - F^2} \cdot m$$

$$m_2 = \frac{\left[\frac{1}{2}(EG - F^2)_2 + FG_1 - GE_2 \right]}{EG - F^2} \cdot m.$$

Recall, finally, that when N has constant curvature K_0 , the intrinsic curvature K_{int} of M and the extrinsic curvature K_{ext} are related by

$$(*) \quad K_{\text{int}} = K_{\text{ext}} + K_0.$$

The first thing we are going to do is to see what the basic lemmas of Chapter 5 give in our more general situation. The main problem is keeping track of the times when the curvature K in the Euclidean case should be replaced by K_{int} and when it should be replaced by K_{ext} .

38. LEMMA. Let M be a surface immersed in a 3-manifold N of constant curvature, and let $p \in M$ be a non-umbilic point. Let $k_1 \geq k_2$ be the two principal curvatures on M and suppose that k_1 has a local maximum at p , and k_2 has a local minimum at p . Then $K_{\text{int}}(p) \leq 0$.

PROOF. The proof is exactly the same as the proof of Lemma 5-1. ♦

39. THEOREM. Let N be a 3-manifold of constant curvature. If M is a compact connected surface in N with constant extrinsic curvature $K_{\text{ext}} \geq 0$ and (constant) intrinsic curvature $K_{\text{int}} > 0$, then all points of M are umbilics.

PROOF. First suppose that $K_{\text{ext}} > 0$. As in the proof of Theorem 5-2, let $k_1 \geq k_2$ be the principal curvatures and let k_1 achieve its maximum at p . Then $k_2 = K_2/k_1$ has its minimum at p . If p were not an umbilic, then by Lemma 38 we would have $K_{\text{int}}(p) \leq 0$, contradicting the hypothesis. So $k_1(p) = k_2(p)$, and, reasoning as in the proof of Theorem 5-2, we see that all points are umbilics.

Next suppose that $K_{\text{ext}} = 0$. Suppose there is a non-umbilic point $p \in M$. Then $0 = k_1(p) \cdot k_2(p)$, but $k_1(p) \neq k_2(p)$, so either $k_1(p) > 0$ or $0 > k_2(p)$, say the first. Let \bar{p} be the point where k_1 takes on its maximum $k_1(\bar{p}) > 0$. Then $k_1 > 0$ in a whole neighborhood of \bar{p} , so $k_2 = 0$ in a whole neighborhood of \bar{p} , and hence k_2 has a local minimum at \bar{p} . Then Lemma 38 gives $K_{\text{int}}(\bar{p}) \leq 0$, a contradiction. ♦

40. THEOREM. Let N be a 3-manifold of constant curvature. Let M be a 2-dimensional immersed submanifold of N with constant extrinsic curvature $K_{\text{ext}} < 0$. Then for every point $p \in M$ there is a diffeomorphism

$$g: (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \rightarrow M$$

$$g(0, 0) = p$$

whose parameter curves are asymptotic curves *parameterized by arclength*.

PROOF. The proof is exactly the same as the first proof of Lemma 5-10. ❖

41. **THEOREM.** Let N be a 3-manifold of constant curvature. Then there is no complete surface M immersed in N with constant extrinsic curvature $K_{\text{ext}} < 0$ and (constant) intrinsic curvature $K_{\text{int}} < 0$.

PROOF. Suppose such a surface M existed. Using Theorem 40, we can repeat the first argument in the (first) proof of Theorem 5-12 *verbatim* and conclude that there is a Tschebyscheff net $f: \mathbb{R}^2 \rightarrow M$. If ω is the angle between the first and second parameter lines, then by Lemma 5-11 we have

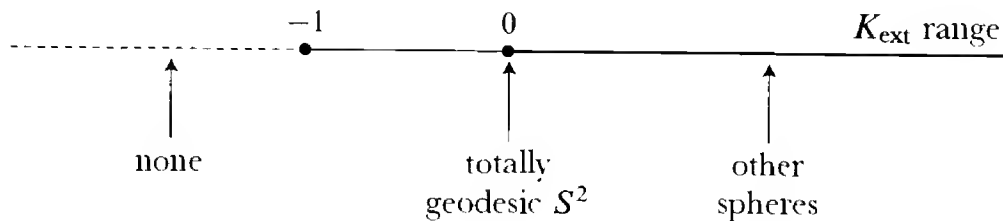
$$\frac{\partial^2 \omega}{\partial s \partial t} = (-K_{\text{int}}) \sin \omega \quad 0 < \omega < \pi,$$

where $-K_{\text{int}}$ is a positive constant. Then part (B) of the proof of Theorem 5-12 shows that there is no such ω . ❖

Now we will begin putting these results together. Take N to be S^3 , with constant curvature 1, and consider the possibilities for complete surfaces in S^3 with constant extrinsic curvature K_{ext} . Since equation (*) now becomes

$$K_{\text{int}} = K_{\text{ext}} + 1,$$

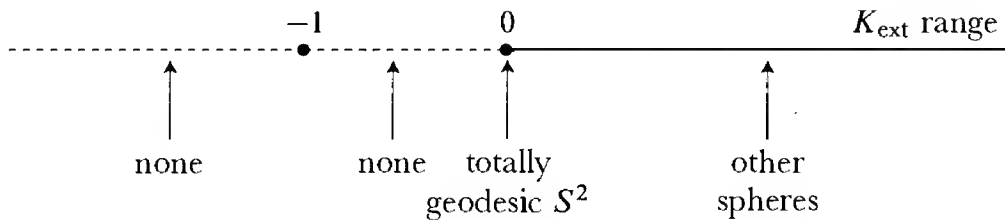
we see that $K_{\text{ext}} < -1 \implies K_{\text{int}} < 0$. So Theorem 41 shows that there are no complete surfaces immersed in S^3 with constant $K_{\text{ext}} < -1$. We also see that $K_{\text{ext}} \geq 0 \implies K_{\text{int}} > 0$, so Theorem 39 and Theorem 27 show that the only compact surfaces in S^3 with constant $K_{\text{ext}} \geq 0$ are spheres (Theorem 8-17 again shows that compactness can be replaced by completeness).



How about the range $-1 \leq K_{\text{ext}} < 0$? First of all we have

42. **PROPOSITION.** There are no complete surfaces M immersed in S^3 with constant K_{ext} satisfying $-1 < K_{\text{ext}} < 0$.

PROOF. The intrinsic curvature of M would satisfy $K_{\text{int}} > 0$, so M would be compact, by Theorem 8-17. We can assume that M is orientable, for otherwise we can look at the orientable 2-fold covering of M , which will also be immersed, with the same K_{ext} . Then M must be homeomorphic to S^2 , by the Gauss-Bonnet Theorem. Since $K_{\text{ext}} < 0$, at every point $p \in S^2$ the principle curvatures $k_1(p), k_2(p)$ have opposite signs. By choosing the vectors pointing in the principal directions which correspond to the positive principal curvature, we would have a continuous choice of 1-dimensional subspaces of S^2_p . But this is impossible (Problem I.9-7). ♦



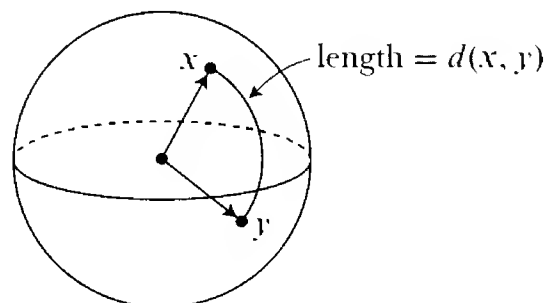
This leaves only the isolated possibility $K_{\text{ext}} = -1$. Oddly enough, there *are* complete surfaces in S^3 with $K_{\text{ext}} = -1$ (equivalently, $K_{\text{int}} = 0$). In fact, for $\rho, \sigma > 0$ with $\rho + \sigma = 1$, the torus

$$\{x \in \mathbb{R}^4 : x_1^2 + x_2^2 = \rho \text{ and } x_3^2 + x_4^2 = \sigma\} \subset S^3$$

is a (flat) product of two circles. Moreover, there is an infinite variety of other complete flat surfaces in S^3 . Such surfaces can be classified, modulo a few sticky details, and we will essentially find the most general way to construct them. The classification actually works even for a piece of a flat surface, but we will deal only with complete surfaces, just to simplify some of the description; this classification is based on the work of Bianchi [1].

It will be necessary to first consider some of the geometry which is special to the manifold S^3 . For two points $x, y \in S^3$, the distance $d(x, y)$ between x and y as elements of S^3 (*not* the Euclidean distance between x and y) is just the radian measure of the angle between x and y . Consequently, we have

$$(1) \quad \cos d(x, y) = \langle x, y \rangle.$$



Now we ask whether there are any isometries $A \in O(4)$ of S^3 with the property that $d(x, A(x))$ is the same for all $x \in S^3$. Such isometries would be the analogues of the translations in \mathbb{R}^n ; notice that S^2 , for example, certainly has no isometries with this property, other than the identity, since every $A \in O(3)$ has a fixed point in S^2 . If $A = (a_{ij})$, then

$$\langle x, Ax \rangle = \sum_{i,j=1}^4 a_{ji} x_j x_i.$$

Taking into account equation (1) we see that we are looking for A with

$$\sum_{i,j=1}^4 a_{ji} x_j x_i = \text{constant} \quad \text{for all } x \in S^4.$$

This implies that

$$\sum_{i,j=1}^4 a_{ji} x_j x_i = (\text{constant}) \cdot \sum_{i=1}^4 x_i^2 \quad \text{for all } x \in \mathbb{R}^4.$$

Regarding this as a polynomial identity in the variables x_1, \dots, x_4 we see that we must have

$$a_{11} = a_{22} = a_{33} = a_{44}, \quad a_{ij} + a_{ji} = 0, \quad i \neq j.$$

Since A is also orthogonal we have

$$(2) \quad 0 = a_{11}a_{12} + a_{21}a_{22} + a_{31}a_{32} + a_{41}a_{42} = a_{31}a_{32} + a_{41}a_{42}$$

as well as

$$a_{11}^2 + a_{21}^2 + a_{31}^2 + a_{41}^2 = a_{12}^2 + a_{22}^2 + a_{32}^2 + a_{42}^2$$

$$\Downarrow$$

$$(3) \quad a_{31}^2 + a_{41}^2 = a_{32}^2 + a_{42}^2.$$

Equation (2) says that the vectors $(a_{31}, a_{41}), (a_{32}, a_{42}) \in \mathbb{R}^2$ are perpendicular, while equation (3) says that they have the same length. It follows that

$$\left. \begin{array}{l} a_{32} = a_{41} \\ a_{42} = -a_{31} \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} a_{32} = -a_{41} \\ a_{42} = +a_{31} \end{array} \right.$$

We thus find two different kinds of A 's with the desired property:

$$\begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix}$$

$$a^2 + b^2 + c^2 + d^2 = 1.$$

The existence of these “translations” in S^3 is directly related to the fact that S^3 is a group, the group of quaternions of norm 1. Recall that the quaternions are \mathbb{R}^4 with the structure of a non-commutative division algebra over \mathbb{R} having unit $1 = (1, 0, 0, 0)$ and elements

$$i = (0, 1, 0, 0), \quad j = (0, 0, 1, 0), \quad k = (0, 0, 0, 1)$$

satisfying

$$\begin{aligned} i \cdot j &= k = -j \cdot i \\ j \cdot k &= i = -k \cdot j \quad \text{and} \quad i \cdot i = j \cdot j = k \cdot k = -1. \\ k \cdot i &= j = -i \cdot k \end{aligned}$$

The norm $|x|$ of a quaternion x satisfies $|xy| = |x| \cdot |y|$, so the quaternions of norm 1 (i.e., S^3) are a non-commutative Lie group. It is easily checked that the two matrices given above are just left and right translation by the quaternion $a + bi + cj + dk \in S^3$. In particular, this shows that the usual Riemannian metric on S^3 is left and right invariant. Moreover, the map

$$a + bi + cj + dk \mapsto \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix}$$

is an isomorphism of S^3 into a subgroup of $O(4)$, namely the subgroup of all left translations by elements of S^3 . It will be convenient to identify S^3 with a subgroup of $O(4)$ by this isomorphism.

We will need the first part of the following general result; the other parts are included for independent interest.

43. THEOREM. Let G be a Lie group with bi-invariant metric $\langle \cdot, \cdot \rangle$. If X, Y, Z, W are left invariant vector fields on G , then

$$(1) \quad \nabla_X Y = \frac{1}{2}[X, Y]$$

- (2) $\langle [X, Y], Z \rangle = \langle X, [Y, Z] \rangle$
 (3) $R(X, Y)Z = -\frac{1}{4}[[X, Y], Z]$
 (4) $\langle R(X, Y)Z, W \rangle = -\frac{1}{4}\langle [X, Y], [Z, W] \rangle$.

PROOF. The integral curves of X are left translates of 1-parameter subgroups (recall the second proof of Corollary I.10-8). Consequently, they are geodesics (Proposition I.10-21). This means that $\nabla_X X = 0$. So

$$0 = \nabla_{X+Y} X + Y = \nabla_X X + \nabla_X Y + \nabla_Y X + \nabla_Y Y = \nabla_X Y + \nabla_Y X.$$

But also

$$\nabla_X Y - \nabla_Y X = [X, Y],$$

which gives (1).

For (2) we note that

$$\begin{aligned} 0 = Y \langle X, Z \rangle &= \langle \nabla_Y X, Z \rangle + \langle X, \nabla_Y Z \rangle \\ &= \frac{1}{2} \langle [Y, X], Z \rangle + \frac{1}{2} \langle X, [Y, Z] \rangle. \end{aligned}$$

For (3) we have

$$\begin{aligned} R(X, Y)Z &= \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X, Y]}Z \\ &= \frac{1}{4}[X, [Y, Z]] - \frac{1}{4}[Y, [X, Z]] - \frac{1}{4}[[X, Y], Z], \end{aligned}$$

which gives the desired result when we apply the Jacobi identity.

Finally, (4) follows from (2) and (3). ♦

Now we want to look at the Lie algebra $\mathcal{L}(S^3)$ of the group S^3 . This is the tangent space of S^3 at $(1, 0, 0, 0)$, and is therefore spanned by the vectors

$$X_1 = (0, 1, 0, 0)$$

$$X_2 = (0, 0, 1, 0)$$

$$X_3 = (0, 0, 0, 1),$$

regarded as tangent vectors at $(1, 0, 0, 0)$. Notice that $X_1 = c'(0)$, where

$$\begin{aligned} c(t) &= (\cos t, \sin t, 0, 0) \in S^3 \\ &= \cos t + (\sin t)i \\ &= \begin{pmatrix} \cos t & -\sin t & 0 & 0 \\ \sin t & \cos t & 0 & 0 \\ 0 & 0 & \cos t & -\sin t \\ 0 & 0 & \sin t & \cos t \end{pmatrix}. \end{aligned}$$

under the identification of S^3 with a subgroup of $O(4)$. Thus X_1 can be identified with

$$c'(0) = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \in \mathfrak{o}(4).$$

Similarly X_2 and X_3 can be identified with

$$\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

A short calculation then shows that

$$(1) \quad [X_1, X_2] = 2X_3, \quad [X_2, X_3] = 2X_1, \quad [X_3, X_1] = 2X_2.$$

If we think of the X_i as vectors in \mathbb{R}^3 , by simply ignoring their first components, then we have

$$X_1 \times X_2 = X_3, \quad X_2 \times X_3 = X_1, \quad X_3 \times X_1 = X_2.$$

Equivalently, this relation holds when we define \times in $S^3_{(1,0,0,0)}$ in terms of the usual inner product and the usual orientation for S^3 . So if \tilde{X}_i is the left invariant vector field on S^3 which extends X_i , then

$$(2) \quad \tilde{X}_1 \times \tilde{X}_2 = \tilde{X}_3, \quad \tilde{X}_2 \times \tilde{X}_3 = \tilde{X}_1, \quad \tilde{X}_3 \times \tilde{X}_1 = \tilde{X}_2,$$

where \times in each tangent space is defined in terms of the usual metric $\langle \cdot, \cdot \rangle$ on S^3 and the usual orientation for S^3 .

Now the theory of curves in S^3 can be given a special development, because we can express all tangent vectors in terms of the left invariant vector fields \tilde{X}_i . Suppose c is a curve in S^3 parameterized by arclength, and let the unit tangent vector $\mathbf{t} = \mathbf{v}_1$ of c be given by

$$(3) \quad \mathbf{t}(s) = \sum_{i=1}^3 f_i(s) \cdot \tilde{X}_i(c(s)),$$

where

$$(4) \quad \sum_i f_i^2 = 1 \implies \sum_i f_i f_i' = 0.$$

As usual, we denote the covariant derivative in our ambient manifold S^3 by ∇' . Then for any vector field $\sum_j h_j(s) \cdot \tilde{X}_j(c(s))$ along c we have

$$\begin{aligned} \frac{D'}{ds} \left[\sum_j h_j(s) \cdot \tilde{X}_j(c(s)) \right] &= \sum_j h_j'(s) \cdot \tilde{X}_j(c(s)) + \sum_j h_j(s) \frac{D'}{ds} \tilde{X}_j(c(s)) \\ &= \sum_j h_j'(s) \cdot \tilde{X}_j(c(s)) \\ &\quad + \sum_j h_j(s) \sum_i f_i(s) \nabla'_{\tilde{X}_i} \tilde{X}_j(c(s)). \end{aligned}$$

Using Theorem 43 to write $\nabla'_{\tilde{X}_i} \tilde{X}_j = \frac{1}{2}[\tilde{X}_i, \tilde{X}_j]$, and computing the brackets from (1), we get

$$(5) \quad \frac{D'}{ds} \left[\sum_j h_j(s) \cdot \tilde{X}_j(c(s)) \right] = \sum_j h_j' \cdot \tilde{X}_j + [(f_2 h_3 - f_3 h_2) \tilde{X}_1 + (f_3 h_1 - f_1 h_3) \tilde{X}_2 + (f_1 h_2 - f_2 h_1) \tilde{X}_3]$$

{all functions evaluated at s , all \tilde{X}_i at $c(s)$ }.

In particular, we have

$$(6) \quad \frac{D' \mathbf{t}(s)}{ds} = \sum_i f_i' \cdot \tilde{X}_i;$$

hence the curvature $\kappa (= \kappa_1)$ is given by

$$(7) \quad \kappa = \sqrt{\sum_i (f_i')^2},$$

and $\mathbf{n} = \mathbf{v}_2$ is given by

$$(8) \quad \mathbf{n} = \frac{\sum_i f_i' \cdot \tilde{X}_i}{\kappa}.$$

Therefore $\mathbf{b} = \mathbf{v}_3$ is given by

$$\begin{aligned} (9) \quad \mathbf{b} = \mathbf{t} \times \mathbf{n} &= \frac{1}{\kappa} \cdot \left(\sum_i f_i \cdot \tilde{X}_i \right) \times \left(\sum_j f_j' \cdot \tilde{X}_j \right) \\ &= \frac{1}{\kappa} \sum_{i,j} f_i f_j' (\tilde{X}_i \times \tilde{X}_j) \\ &= \frac{1}{\kappa} [(f_2 f_3' - f_3 f_2') \tilde{X}_1 + (f_3 f_1' - f_1 f_3') \tilde{X}_2 + (f_1 f_2' - f_2 f_1') \tilde{X}_3] \\ &\quad \text{by (2)} \\ &= \frac{1}{\kappa} \sum_i g_i \cdot \tilde{X}_i, \quad \text{say.} \end{aligned}$$

Now we have

$$\begin{aligned}
 \frac{D'\mathbf{b}(s)}{ds} &= \frac{1}{\kappa} \frac{D'}{ds} \left(\sum_i g_i \cdot \tilde{X}_i \right) - \frac{\kappa'}{\kappa^2} \sum_i g_i \cdot \tilde{X}_i \\
 &= \frac{1}{\kappa} \left[(f_2 g_3 - g_2 f_3) \tilde{X}_1 + \cdots + \sum_i g_i' \tilde{X}_i \right] - \frac{\kappa'}{\kappa^2} \sum_i g_i \tilde{X}_i \quad \text{by (5)} \\
 &= \frac{1}{\kappa} [(f_2 g_3 - g_2 f_3) \tilde{X}_1 + \cdots] + \sum_i \left(\frac{g_i}{\kappa} \right)' \tilde{X}_i.
 \end{aligned}$$

But

$$\begin{aligned}
 f_2 g_3 - g_2 f_3 &= f_2(f_1 f_2' - f_2 f_1') - f_3(f_3 f_1' - f_1 f_3') \quad \text{by (9)} \\
 &= f_1(f_2 f_2' + f_3 f_3') - f_1'(f_2^2 + f_3^2) \\
 &= f_1(-f_1 f_1') - f_1'(1 - f_1^2) \quad \text{by (4)} \\
 &= -f_1',
 \end{aligned}$$

and similarly for the other terms. Hence we obtain

$$\begin{aligned}
 (10) \quad \frac{D'\mathbf{b}(s)}{ds} &= \frac{-\sum_i f_i' \cdot \tilde{X}_i}{\kappa} + \sum_i \left(\frac{g_i}{\kappa} \right)' \cdot \tilde{X}_i \\
 &= -\mathbf{n} + \sum_i \left(\frac{g_i}{\kappa} \right)' \cdot \tilde{X}_i \quad \text{by (8)}.
 \end{aligned}$$

We therefore have the rather remarkable, and for us very important

44. THEOREM. If c is a curve in S^3 whose torsion τ ($= \kappa_2$) satisfies $\tau = 1$ everywhere, then \mathbf{b} is left invariant along c , that is,

$$b(s) = L_{c(s)c(0)^{-1}*} b(0).$$

If c has torsion $\tau = -1$ everywhere, then \mathbf{b} is right invariant along c .

PROOF. The Serret-Frenet formulas give

$$\frac{D\mathbf{b}(s)}{ds} = -\tau \mathbf{n}.$$

So $\tau = 1$ implies that $(g_i/\kappa)' = 0$, and hence that g_i/κ is constant. But equation (9) shows that g_i/κ are the components of \mathbf{b} with respect to the left invariant vector fields \tilde{X}_i .

To deduce the second part of the theorem, consider the map $f(x) = x^{-1}$ of S^3 into itself. This map reverses 1-parameter subgroups through $(1, 0, 0, 0)$, so $f_*: \mathcal{L}(S^3) \rightarrow \mathcal{L}(S^3)$ is multiplication by -1 . This shows that f is orientation reversing. It follows that the binormal of the curve $f \circ c$ is $-f_*\mathbf{b}$. Thus $f \circ c$ has $\tau = 1$ if and only if c has $\tau = -1$. ♦

Finally we are ready to consider connected immersed surfaces M in S^3 with $K_{\text{ext}} = -1$, and hence $K_{\text{int}} = 0$. We consider only oriented M ; non-orientable surfaces may then be analyzed by considering the 2-fold oriented covering of M . Since $K_{\text{ext}} < 0$, there are 2 distinct asymptotic directions at each point. The argument in the (first) proof of Theorem 5-12, in conjunction with Theorem 40, again shows that there is a Tsychebyscheff net $f: \mathbb{R}^2 \rightarrow M$. It is not hard to see that f is actually onto M (by essentially the argument used in the second proof of Theorem 5-12; for this part, it is not necessary that the ϕ_s be defined for all $s \in \mathbb{R}$, and simple-connectivity is irrelevant). The metric $I_f = f^*\langle \cdot, \cdot \rangle$ on \mathbb{R}^2 is then

$$I_f = f^*\langle \cdot, \cdot \rangle = ds \otimes ds + \cos \omega [ds \otimes dt + dt \otimes ds] + dt \otimes dt,$$

where ω is the oriented angle between the first and second parameter curves.

Now consider the curve $c(s) = (s, t)$ in \mathbb{R}^2 , which is an arclength parameterized curve for the metric I_f . Its tangent vector $c'(s) = \partial/\partial s$ is a unit vector for the metric I_f . If D/ds temporarily denotes the covariant derivative determined by the metric I_f , then from the formula on pg. II.232 we compute that

$$\frac{Dc'(s)}{ds} = \frac{\frac{\partial \omega}{\partial s}}{\sin \omega} \cdot \left[\cos \omega \cdot \frac{\partial}{\partial s} - \frac{\partial}{\partial t} \right].$$

If $\left(\frac{\partial}{\partial s}\right)^\perp$ is the unique vector field with $\frac{\partial}{\partial s}, \left(\frac{\partial}{\partial s}\right)^\perp$ orthonormal for the metric I_f and $\frac{\partial}{\partial s}, \left(\frac{\partial}{\partial s}\right)^\perp$ positively oriented, then

$$\frac{\partial}{\partial t} = \cos \omega \cdot \frac{\partial}{\partial s} + \sin \omega \cdot \left(\frac{\partial}{\partial s}\right)^\perp,$$

so we find that

$$\frac{Dc'(s)}{ds} = -\frac{\frac{\partial \omega}{\partial s}}{\sin \omega} \cdot \left(\frac{\partial}{\partial s}\right)^\perp.$$

Equivalently, if \mathbf{t} denotes the (unit) tangent vector to the parameter curve $s \mapsto f(s, t)$ in M , and D/ds now denotes the covariant derivative in M , then

$$\frac{D\mathbf{t}}{ds} = -\frac{\frac{\partial \omega}{\partial s}}{\sin \omega} \cdot \mathbf{u},$$

where \mathbf{u} is the unique tangent vector field along $s \mapsto f(s, t)$ with \mathbf{t}, \mathbf{u} orthonormal and (\mathbf{t}, \mathbf{u}) positively oriented. But $s \mapsto f(s, t)$ is an asymptotic curve, so the

covariant derivative $D\mathbf{t}/ds$ in M is the same as the covariant derivative $D'\mathbf{t}/ds$ in S^3 (recall the equivalences on pg. III.196). So we have

$$(1) \quad \frac{D'\mathbf{t}}{ds} = -\frac{\partial\omega}{\partial s} \cdot \mathbf{u}.$$

This shows that

$$\begin{aligned} \mathbf{u} &= \text{normal } \mathbf{n} \text{ to the curve } s \mapsto f(s, t) \\ \left| \frac{\partial\omega}{\partial s}(s, t) \right| &= \text{curvature } \kappa(s) \text{ of the curve } s \mapsto f(s, t). \end{aligned}$$

On the other hand, Lemma 5-11 shows that ω satisfies

$$\frac{\partial^2\omega}{\partial s\partial t} = 0,$$

which implies that there are functions S and T with

$$\omega(s, t) = S(s) + T(t),$$

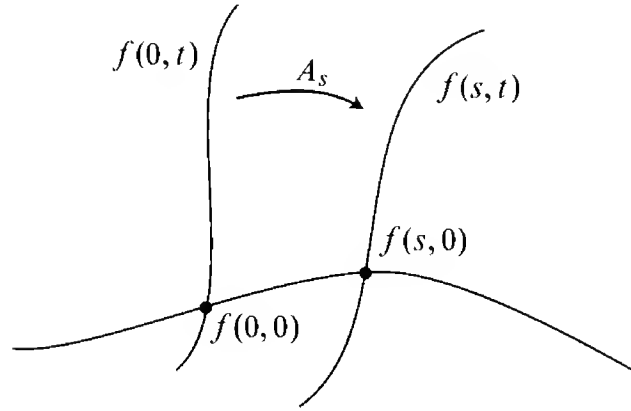
so that

$$\frac{\partial\omega}{\partial s}(s, t) = S'(s) \quad \text{and} \quad \frac{\partial\omega}{\partial t}(s, t) = T'(t).$$

Thus the arclength parameterized curves $s \mapsto f(s, t)$ all have the *same* curvature functions $\kappa(s) = |S'(s)|$. Similarly, all curves $t \mapsto f(s, t)$ have the same curvature functions $|T'(t)|$.

But even more is true. For the Beltrami-Enneper Theorem (Theorem 36) tells us that the torsion τ of the asymptotic curves $s \mapsto f(s, t)$ and $t \mapsto f(s, t)$ satisfies $\tau^2 = 1$ at points where $\kappa \neq 0$, and that the two asymptotic curves through a point have torsions of opposite signs if they both have $\kappa \neq 0$ at that point. We will first assume that for both sets of parameter curves κ is never 0. Then one set of parameter curves must have $\tau = 1$ everywhere, and the other set must have $\tau = -1$ everywhere. For definiteness, say that the curves $s \mapsto f(s, t)$ have $\tau = 1$. We now see that *all* curves $s \mapsto f(s, t)$ are congruent, and similarly *all* curves $t \mapsto f(s, t)$ are congruent.

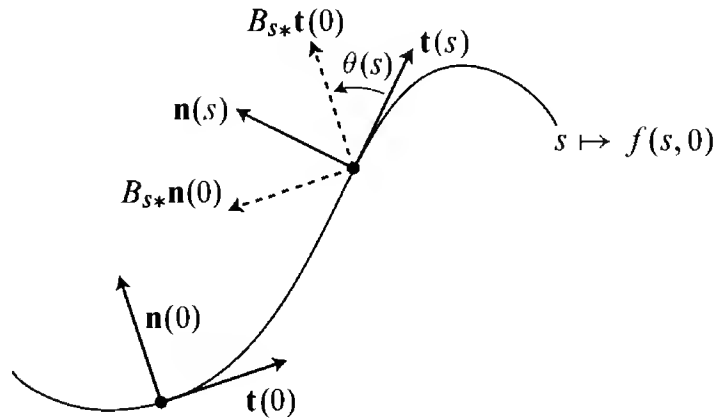
Let A_s be the unique isometry of S^3 onto itself with $A_s(f(0, t)) = f(s, t)$ for all t . Under the family of isometries $\{A_s\}$, each point $f(0, t)$ moves along the arclength parameterized curve $s \mapsto f(s, t)$. This strongly suggests that all the A_s are actually translations. In fact, we claim that all A_s are left translations.



To prove this, we consider the family of left translations $\{B_s\} = \{L_{f(s,0)f(0,0)^{-1}}\}$ which take $f(0, 0)$ to $f(s, 0)$. According to Theorem 44, B_{s*} takes the binormal $\mathbf{b}(0)$ of $s \mapsto f(s, 0)$ at $s = 0$ into the binormal $\mathbf{b}(s)$ at s . Consequently, B_{s*} takes the osculating plane of this curve at 0 into the osculating plane at s . Hence we can write

$$\mathbf{t}(s) = \cos \theta(s) \cdot B_{s*}\mathbf{t}(0) - \sin \theta(s) \cdot B_{s*}\mathbf{n}(0),$$

where $\theta(s)$ is the oriented angle from $\mathbf{t}(s)$ to $B_{s*}\mathbf{t}(0)$. It is easy to compute



$D'\mathbf{t}/ds$ in terms of θ : For simplicity, and without loss of generality, we assume that $f(0, 0) = 1 \in S^3$, and that $\mathbf{t}(0)$ and $\mathbf{n}(0)$ are $X_1, X_2 \in \mathcal{L}(S^3)$. Then the functions f_i in equation (3) on page 98 are just

$$f_1 = \cos \theta, \quad f_2 = -\sin \theta, \quad f_3 = 0,$$

so equation (6) on page 99 gives

$$\begin{aligned} \frac{D'\mathbf{t}}{ds} &= -\theta'(s)[\sin \theta(s) \cdot B_{s*}\mathbf{t}(0) + \cos \theta \cdot B_{s*}\mathbf{n}(0)] \\ &= -\theta'(s) \cdot \mathbf{n}(s). \end{aligned}$$

Comparing with equation (1) on page 102, we see that $\theta' = S'$; since $\theta(0) = 0$, we find that

$$S(s) = \theta(s) + S(0).$$

From this we easily see that

B_{s*} takes the tangent vector to the curve $t \mapsto f(0, t)$ at $t = 0$
to the tangent vector to the curve $t \mapsto f(s, t)$ at $t = 0$.

Moreover, these curves are asymptotic curves, so their osculating planes at $t = 0$ coincide with the osculating planes of the asymptotic curve $s \mapsto f(s, 0)$ at 0 and s , respectively. Thus their binormals at $t = 0$ are the binormals $\mathbf{b}(0)$ and $\mathbf{b}(s)$ of the curve $s \mapsto f(s, t)$. Hence

B_{s*} takes the binormal to the curve $t \mapsto f(0, t)$ at $t = 0$
to the binormal to the curve $t \mapsto f(s, t)$ at $t = 0$.

These two facts show that B_s must be the isometry A_s . So A_s is indeed a left translation.

If we write $c(s) = f(s, 0)$ and $\gamma(t) = f(0, t)$, we thus see that our surface M can be written as a collection of left translates of γ ,

$$M = \{[c(s) \cdot c(0)^{-1}] \cdot \gamma(t)\}.$$

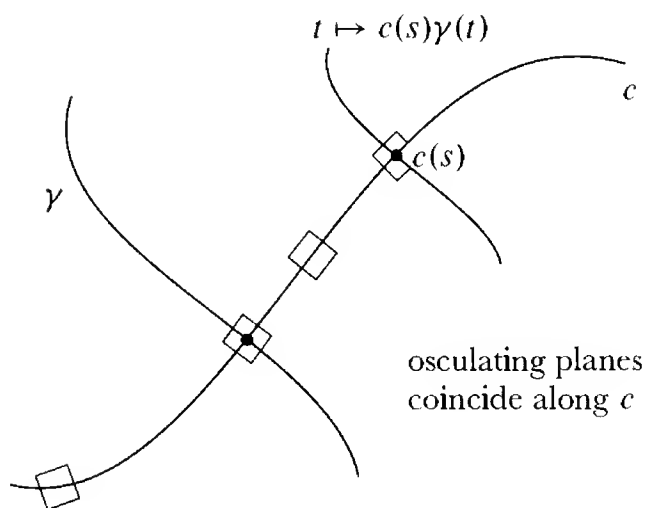
Notice that this can equally well be written as a collection of right translates of c ,

$$\begin{aligned} M &= \{c(s) \cdot c(0)^{-1} \cdot \gamma(t)\} = \{c(s) \cdot \gamma(0)^{-1} \cdot \gamma(t)\} \\ &= \{c(s) \cdot [\gamma(0)^{-1} \cdot \gamma(t)]\}; \end{aligned}$$

naturally we could have also deduced this description directly, by considering the isometries of the curves $s \mapsto f(s, t)$, and applying the second part of Theorem 44.

Conversely, suppose we have any two curves c and γ with torsions $\tau = 1$ and $\tau = -1$, respectively. Suppose, moreover, that they are placed so that $c(0) = \gamma(0)$ and so that their osculating planes at 0 coincide. For simplicity, also assume that $c(0) = \gamma(0) = 1 \in S^3$. Then c and γ will not be tangent at 0, and we can consider the surface

$$M = \{c(s) \cdot \gamma(t)\}.$$

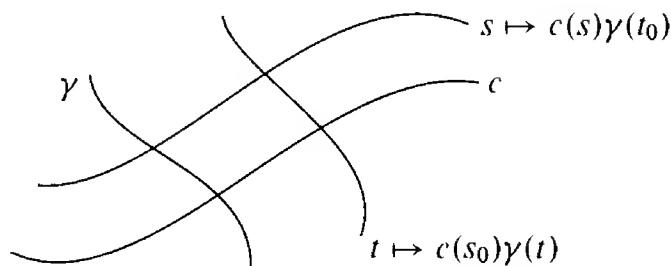


Applying Theorem 44 first to the curve c with $\tau = 1$, we find that the osculating plane of c at s coincides with the osculating plane of the curve $t \mapsto c(s) \cdot \gamma(t)$ at $t = 0$; hence these osculating planes coincide with the tangent space of M at $c(s)$. Now applying Theorem 44 to the curves $t \mapsto c(s) \cdot \gamma(t)$, all with torsions $\tau = -1$, we find that the tangent space of M at any point $c(s_0) \cdot \gamma(t_0)$ coincides with the osculating planes of the parameter curves $s \mapsto c(s) \cdot \gamma(t_0)$ at $s = s_0$ and $t \mapsto c(s_0) \cdot \gamma(t)$ at $t = t_0$. Thus these parameter curves are asymptotic curves. So the Beltrami-Enneper Theorem shows that M has $K_{\text{ext}} = -1$.

We can also consider the case where c has torsion $\tau = 1$, but γ is a geodesic, and hence does not have a torsion defined anywhere. The first part of our argument still shows that the tangent space of M at points $c(s)$ coincides with the osculating plane of c at s . In other words,

$$(1) \quad \frac{D'c'(s)}{ds} \text{ is a linear combination of } c'(s) \text{ and } L_{c(s)*}\gamma'(0).$$

To show that the tangent space of M at $c(s_0) \cdot \gamma(t_0)$ coincides with the osculating plane of $s \mapsto c(s) \cdot \gamma(t_0)$ at $s = s_0$, we must show that



$\left. \frac{D'}{ds} \right|_{s=s_0} R_{\gamma(t_0)*} c'(s)$ is a linear combination of $R_{\gamma(t_0)*} c'(s_0)$ and $L_{c(s_0)*} \gamma'(t_0)$.

Now we have

$$\begin{aligned} \left. \frac{D'}{ds} \right|_{s=s_0} R_{\gamma(t_0)*} c'(s) &= R_{\gamma(t_0)*} \left. \frac{D' c'(s)}{ds} \right|_{s=s_0} \\ &= \text{a linear combination of } R_{\gamma(t_0)*} c'(s_0) \\ &\quad \text{and } R_{\gamma(t_0)*} L_{c(s_0)*} \gamma'(0), \quad \text{by (1).} \end{aligned}$$

So it suffices to observe that

$$\begin{aligned} R_{\gamma(t_0)*} L_{c(s_0)*} \gamma'(0) &= L_{c(s_0)*} R_{\gamma(t_0)*} \gamma'(0) \\ &= L_{c(s_0)*} \gamma'(t_0), \end{aligned}$$

since the geodesic γ through $1 \in S^3$ is a 1-parameter subgroup, and hence the integral curve of a right invariant vector field (recall again the second proof of Corollary I.10-8; although this proof deals with left invariant vector fields, it works just as well for right invariant vector fields). So our surface $M = \{c(s) \cdot \gamma(t)\}$ again has $K_{\text{ext}} = -1$.

Finally,* suppose that c and γ are *both* (distinct) geodesics through $1 \in S^3$. Then the surface $M = \{c(s) \cdot \gamma(t)\}$ still has $K_{\text{ext}} = -1$, or $K_{\text{int}} = 0$. To see this, we consider the parameter curves

$$\begin{array}{lll} s \mapsto c(s) \cdot \gamma(t_0) & \text{with tangent vectors} & R_{\gamma(t_0)*} c'(s_0) \\ t \mapsto c(s_0) \cdot \gamma(t) & & L_{c(s_0)*} \gamma'(t_0). \end{array}$$

We note that

$$\begin{aligned} \langle R_{\gamma(t_0)*} c'(s_0), L_{c(s_0)*} \gamma'(t_0) \rangle &= \langle R_{\gamma(t_0)*} L_{c(s_0)*} c'(0), L_{c(s_0)*} R_{\gamma(t_0)*} \gamma'(0) \rangle \\ &= \langle R_{\gamma(t_0)*} L_{c(s_0)*} c'(0), R_{\gamma(t_0)*} L_{c(s_0)*} \gamma'(0) \rangle \\ &= \langle c'(0), \gamma'(0) \rangle. \end{aligned}$$

Thus our surface has two families of geodesics intersecting at a constant angle, so it is flat by Proposition 4-6. In particular, the flat torus

$$\left\{ x \in \mathbb{R}^4 : x_1^2 + x_2^2 = \frac{1}{2} \text{ and } x_3^2 + x_4^2 = \frac{1}{2} \right\}$$

*We will not consider the case where our asymptotic curves have curvature $\kappa(s) = 0$ for only certain s . The truly fanatical reader may wish to investigate this situation further.

is of this form. It is generated by the two geodesics

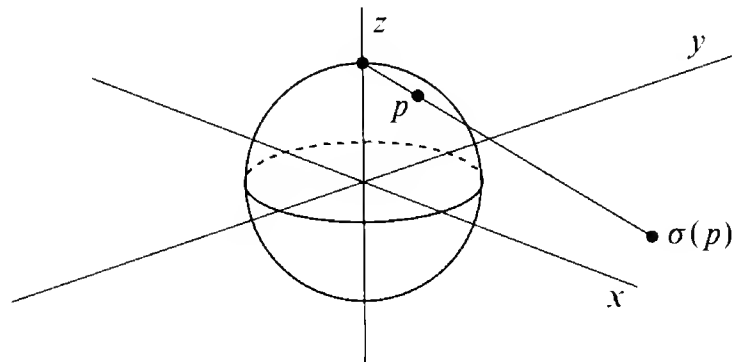
$$\left\{ \frac{1}{\sqrt{2}}(\cos \theta, \sin \theta, \cos \theta, \sin \theta) \right\}$$

$$\left\{ \frac{1}{2}(\cos \phi + \sin \phi, \cos \phi - \sin \phi, -\cos \phi - \sin \phi, -\cos \phi + \sin \phi) \right\},$$

of which the second lies in the plane spanned by $(1, 1, -1, -1), (1, -1, -1, 1)$ and the first in the orthogonal complement.

We now have a very general way of describing surfaces M in S^3 with $K_{\text{ext}} = -1$; we can take any “translation surface” $\{c(s) \cdot \gamma(t)\}$, where c and γ are curves of torsion 1 and -1 with $c(0) = \gamma(0) = 1 \in S^3$ and common osculating planes at 1. Since the curves c and γ are otherwise arbitrary, there are clearly a great number of such surfaces. We will describe some features of these surfaces in a little greater detail, and then indicate some open questions.

It will be very useful to introduce a famous creature of algebraic topology, the Hopf map $h: S^3 \rightarrow S^2$, which is defined as follows. We regard S^2 as the one-point compactification $\mathbb{C} \cup \{\infty\}$ of the complex numbers; the specific identification of S^2 and $\mathbb{C} \cup \{\infty\}$ will be given by means of stereographic projection, together with the identification of the north pole of S^2 with ∞ . However, we will use a slightly different version of stereographic projection. We now regard S^2 as the standard unit sphere $\{p \in \mathbb{R}^3 : |p| = 1\}$, and map a point $p \in S^2 - \{(0, 0, 1)\}$ into the intersection $\sigma(p)$ of the (x, y) -plane with the straight line between $(0, 0, 1)$ and p . It is easy to check that for our new σ we have



$$\sigma(a, b, c) = \left(\frac{a}{1-c}, \frac{b}{1-c} \right)$$

$$\sigma^{-1}(x, y) = \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1} \right).$$

It is not hard to see that $S^2 = \mathbb{C} \cup \{\infty\}$ has a C^∞ atlas consisting of two maps

$$\begin{aligned} f_1: \mathbb{C} &\rightarrow \mathbb{C} \\ f_2: \mathbb{C} - \{0\} \cup \{\infty\} &\rightarrow \mathbb{C} \end{aligned}$$

with $f_1 = \text{identity}$ and

$$f_2(z) = \begin{cases} \frac{1}{z}, & z \neq \infty \\ 0, & z = \infty. \end{cases}$$

We consider S^3 as

$$\{(z_1, z_2) \in \mathbb{C} \times \mathbb{C} : |z_1|^2 + |z_2|^2 = 1\}.$$

Then $h: S^3 \rightarrow S^2$ is defined by

$$h(z_1, z_2) = \frac{z_1}{z_2},$$

where “ z_1/z_2 ” = ∞ if $z_2 = 0$. This map is clearly C^∞ on the set where $z_2 \neq 0$, and also on the set where $z_1 \neq 0$, since we then have

$$\begin{aligned} f_2 \circ h(z_1, z_2) &= \begin{cases} f_2\left(\frac{z_1}{z_2}\right), & z_2 \neq 0 \\ f_2(\infty), & z_2 = 0 \end{cases} = \begin{cases} \frac{z_2}{z_1}, & z_2 \neq 0 \\ 0, & z_2 = 0 \end{cases} \\ &= \frac{z_2}{z_1}. \end{aligned}$$

The inverse image $h^{-1}(z_0)$ of any point $z_0 \in \mathbb{C}$ is

$$h^{-1}(z_0) = \{(z_1, z_2) \in S^3 : z_1 = z_0 z_2\}.$$

If $z_j = x_j + iy_j$ for $j = 0, 1, 2$, this can be written as

$$h^{-1}(z_0) = \{(x_1, y_1, x_2, y_2) \in S^3 : x_1 = x_0 x_2 - y_0 y_2 \text{ and } y_1 = x_0 y_2 + x_2 y_0\},$$

which is the intersection of S^3 with two hyperplanes through the origin. So $h^{-1}(z_0)$ is a great circle. Moreover,

$$h^{-1}(\infty) = \{(z_1, z_2) \in S^3 : z_2 = 0\}$$

is also a great circle.

Now we need to know what the orthogonal maps $A: S^2 \rightarrow S^2$ look like when we consider them as maps $\mathbb{C} \cup \{\infty\} \rightarrow \mathbb{C} \cup \{\infty\}$. Elementary complex analysis tells us that they must be maps of the form

$$f(z) = (az + b)/(cz + d),$$

for they must be one-one and have at most one pole, of order ≤ 1 (we can also use Problem 4-11 to reach the same conclusion). Some further calculations (Problem 21) show that these maps, when normalized to have $ad - bc = 1$, correspond to orthogonal maps if and only if

$$\begin{aligned} |a|^2 + |c|^2 &= 1 \\ |b|^2 + |d|^2 &= 1 \end{aligned} \quad a\bar{b} = -c\bar{d}.$$

On the other hand, if these conditions are satisfied, then the map

$$g(z_1, z_2) = (az_1 + bz_2, cz_1 + dz_2)$$

is easily seen to be an isometry of $S^3 \subset \mathbb{C} \times \mathbb{C}$. Now for any set $X \subset S^2$ we have

$$\begin{aligned} (z_1, z_2) \in h^{-1}(f^{-1}(X)) &\iff (z_1, z_2) \in S^3 \text{ and } \frac{z_1}{z_2} \in f^{-1}(X) \\ &\iff (z_1, z_2) \in S^3 \text{ and } \frac{a\frac{z_1}{z_2} + b}{c\frac{z_1}{z_2} + d} \in X \\ &\iff (z_1, z_2) \in S^3 \text{ and } \frac{az_1 + bz_2}{cz_1 + dz_2} \in X \\ &\iff (z_1, z_2) \in S^3 \text{ and } h(g(z_1, z_2)) \in X. \end{aligned}$$

Thus

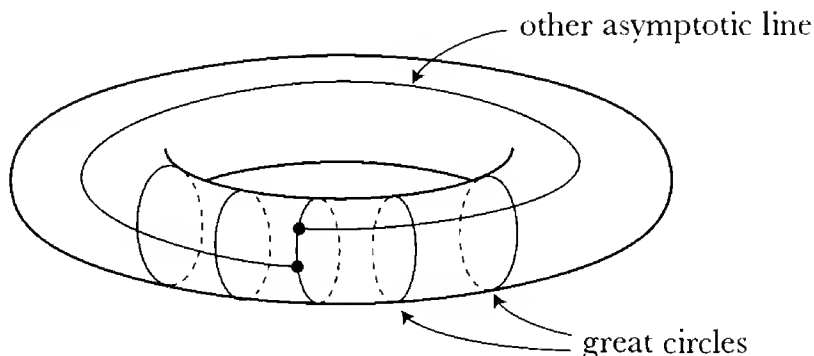
$$h^{-1}(f^{-1}(X)) = g^{-1}(h^{-1}(X)).$$

In other words, if we want to know what $h^{-1}(X) \subset S^3$ looks like, up to an isometry of S^3 , we can replace $X \subset S^2$ by any set related to X by an isometry of S^2 . In particular, to find $h^{-1}(\Sigma)$ for $\Sigma \subset S^2$ a circle, we can assume that Σ is parallel to the (x, y) -plane, so that the stereographic projection of Σ in \mathbb{C} is just a circle $\{z : |z| = R\}$. Then

$$\begin{aligned} h^{-1}(\{z : |z| = R\}) &= \left\{ (z_1, z_2) : |z_1|^2 + |z_2|^2 = 1 \text{ and } \left| \frac{z_1}{z_2} \right| = R \right\} \\ &= \left\{ (z_1, z_2) : |z_1| = \frac{R}{\sqrt{1+R^2}} \text{ and } |z_2| = \frac{1}{\sqrt{1+R^2}} \right\}, \end{aligned}$$

which is just a product torus. This shows that all product tori in S^3 are made up of a family of great circles, which are consequently asymptotic curves. When $R \neq 1$, the other asymptotic curves are not great circles. If they begin at one

point of a great circle they will generally return to a different point of this great circle. This shows how a translation surface $\{c(s) \cdot \gamma(t)\}$ can be compact even though the curve c or γ may not be closed.

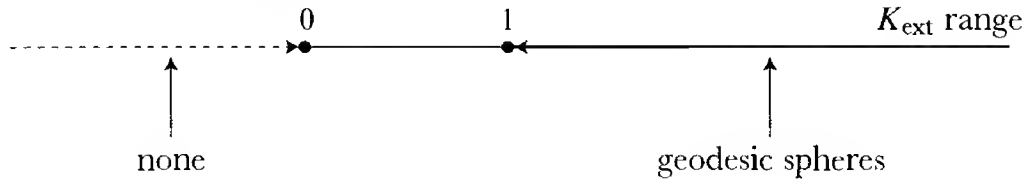


Now let c be any immersed curve S^3 . We claim that the surface $h^{-1}(c)$ has $K_{\text{ext}} = -1$ everywhere. In fact, for any s_0 , we can consider the osculating circle $\Sigma \subset S^2$ of c at s_0 (in other words, Σ is the circle in S^2 which is tangent to c at s_0 and whose curvature, as a curve in S^2 , is the same as the curvature $\kappa(s_0)$ of c at s_0). Then Σ and c agree up to second order at $c(s_0)$, so $h^{-1}(\Sigma)$ and $h^{-1}(c)$ agree up to second order on the whole great circle $h^{-1}(c(s_0))$; since $h^{-1}(\Sigma)$ is a flat torus, with $K_{\text{ext}} = -1$ everywhere, $h^{-1}(c)$ must also have $K_{\text{ext}} = -1$ everywhere. Taking c to be an imbedded closed curve in S^2 , we obtain an imbedded surface $h^{-1}(c)$ in S^3 , with $K_{\text{ext}} = -1$, which is homeomorphic to a torus, but generally not a product torus. A non-geodesic asymptotic curve in $h^{-1}(c)$ will be a curve \tilde{c} with $h \circ \tilde{c} = c$; it would be interesting (and probably very difficult) to determine for precisely which curves c this curve \tilde{c} is closed. In this connection, we point out that there are certainly some closed curves in S^3 of constant torsion $\tau = 1$. In fact, just as cylindrical helices in \mathbb{R}^3 have constant torsion, the helices on product tori in S^3 are easily seen to have constant torsion, and in the latter case we can arrange for the helices to be closed. I do not know whether there are closed curves c and γ in S^3 of torsion $\tau = +1$ and $\tau = -1$ such that the translation surface $\{c(s) \cdot \gamma(t)\}$ is an *imbedded* torus (the helices on product tori give only immersed tori). Nor do I know the answer to the following problem, which seems quite hard: are there one-one curves c and γ in S^3 of torsion $\tau = +1$ and $\tau = -1$ such that the translation surface $\{c(s) \cdot \gamma(t)\}$ is a *one-one* map into S^3 ? Finally, one could try to analyze the non-orientable complete surfaces in S^3 with $K_{\text{ext}} = -1$.

Now we consider the case $N = H^3$, with constant curvature -1 , so that (*) becomes

$$K_{\text{int}} = K_{\text{ext}} - 1.$$

First we see that $K_{\text{ext}} < 0 \implies K_{\text{int}} < 0$, so Theorem 41 shows that there are no complete surfaces immersed in H^3 with constant $K_{\text{ext}} < 0$. Since we also have $K_{\text{ext}} > 1 \implies K_{\text{int}} > 0$, Theorem 39 implies that a complete surface immersed in H^3 with constant $K_{\text{ext}} > 1$ is all-umbilic; since $K_{\text{int}} > 0$, it must actually be a geodesic sphere.



In the range $0 \leq K_{\text{ext}} \leq 1$ we have at least the totally geodesic spheres, the equidistant surfaces, and the horospheres, but we will find other examples also.

We consider first the upper range $K_{\text{ext}} = 1 \implies K_{\text{int}} = 0$. By considering the universal covering space of our immersed surface M with $K_{\text{int}} = 0$ we can assume that M is simply-connected. Thus M , with the induced metric, is isometric to \mathbb{R}^2 with its usual metric. Equivalently, we are considering *isometric* immersions $f: \mathbb{R}^2 \rightarrow H^3$, where \mathbb{R}^2 has its usual metric $dx \otimes dx + dy \otimes dy$, and H^3 has the metric $\langle \cdot, \cdot \rangle$ of constant curvature -1 . Let l_{ij} be the coefficients of the second fundamental form Π_f . In Gauss' equation,

$$\begin{aligned} & \langle s(X, Z), s(Y, W) \rangle - \langle s(Y, Z), s(X, W) \rangle \\ &= \langle R'(X, Y)Z, W \rangle - \langle R(X, Y)Z, W \rangle \\ &= -[\langle X, W \rangle \cdot \langle Y, Z \rangle - \langle X, Z \rangle \cdot \langle Y, W \rangle] - \langle R(X, Y)Z, W \rangle, \end{aligned}$$

we choose $X = Z = \partial/\partial x$ and $Y = W = \partial/\partial y$, to obtain

$$(1) \quad l_{11}l_{22} - (l_{12})^2 = 1.$$

In the Codazzi-Mainardi equations,

$$\begin{aligned} 0 &= (\nabla_X \Pi)(Y, Z) - (\nabla_Y \Pi)(X, Z) \\ &= X(\Pi(Y, Z)) - Y(\Pi(X, Z)) - \dots + \dots, \end{aligned}$$

we take $X = \partial/\partial x$ and $Y = \partial/\partial y$, and then $Z = \partial/\partial x$ or $\partial/\partial y$ to obtain

$$(2) \quad \frac{\partial l_{12}}{\partial x} = \frac{\partial l_{11}}{\partial y}, \quad \frac{\partial l_{22}}{\partial x} = \frac{\partial l_{12}}{\partial y}.$$

These equations imply that there are functions $\alpha, \beta: \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$\begin{array}{ll} (a) \quad \frac{\partial \alpha}{\partial y} = l_{12} & (c) \quad \frac{\partial \beta}{\partial y} = l_{22} \\ (b) \quad \frac{\partial \alpha}{\partial x} = l_{11} & (d) \quad \frac{\partial \beta}{\partial x} = l_{12}. \end{array} \quad \text{and}$$

Then (a) and (d) imply that there is a function $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$\frac{\partial \phi}{\partial x} = \alpha \quad \text{and} \quad \frac{\partial \phi}{\partial y} = \beta.$$

Together with (b) and (c) we thus have

$$(3) \quad \frac{\partial^2 \phi}{\partial x^2} = l_{11}, \quad \frac{\partial^2 \phi}{\partial x \partial y} = l_{12}, \quad \frac{\partial^2 \phi}{\partial y^2} = l_{22}.$$

Thus equation (1) yields

$$(*) \quad \frac{\partial^2 \phi}{\partial x^2} \frac{\partial^2 \phi}{\partial y^2} - \left(\frac{\partial^2 \phi}{\partial x \partial y} \right)^2 = 1.$$

We now appeal to a strange result which is usually used in a completely different context (see Chapter 9):

45. THEOREM (JÖRGENS). If $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a function on the whole plane satisfying

$$(*) \quad \frac{\partial^2 \phi}{\partial x^2} \frac{\partial^2 \phi}{\partial y^2} - \left(\frac{\partial^2 \phi}{\partial x \partial y} \right)^2 = 1,$$

then ϕ is a quadratic polynomial in x and y .

PROOF. We adopt the abbreviations

$$\begin{aligned} p &= \frac{\partial \phi}{\partial x}, & q &= \frac{\partial \phi}{\partial y} \\ r &= \frac{\partial^2 \phi}{\partial x^2}, & s &= \frac{\partial^2 \phi}{\partial x \partial y}, & t &= \frac{\partial^2 \phi}{\partial y^2}, \end{aligned}$$

so that our equation reads

$$(*) \quad rt - s^2 = 1.$$

This implies that $rt > 0$, so that r and t have the same sign. We can assume that $r, t > 0$ everywhere, by replacing ϕ by $-\phi$ if necessary.

For fixed (x_0, y_0) and (x_1, y_1) , consider the function

$$h(\tau) = \phi(x_0 + \tau(x_1 - x_0), y_0 + \tau(y_1 - y_0)).$$

We have

$$\begin{aligned} h'(\tau) &= (x_1 - x_0)p + (y_1 - y_0)q, \\ h''(\tau) &= (x_1 - x_0)^2 r + 2(x_1 - x_0)(y_1 - y_0)s + (y_1 - y_0)^2 t, \end{aligned}$$

where p, q, r, s, t are evaluated at $(x_0 + \tau(x_1 - x_0), y_0 + \tau(y_1 - y_0))$. If $x_1 = x_0$, then $h''(\tau) = (y_1 - y_0)^2 t \geq 0$. If $x_1 \neq x_0$, then

$$h''(\tau) = (x_1 - x_0)^2 \left[r - 2 \left(\frac{y_1 - y_0}{x_1 - x_0} \right) s + \left(\frac{y_1 - y_0}{x_1 - x_0} \right)^2 t \right].$$

The term in brackets is a quadratic polynomial in $(y_1 - y_0)/(x_1 - x_0)$ with discriminant $4s^2 - 4rt < 0$, by (*), so it is always positive. Thus we always have $h''(\tau) \geq 0$. This implies that

$$h'(1) \geq h'(0),$$

and thus

$$(1) \quad (x_1 - x_0)(p_1 - p_0) + (y_1 - y_0)(q_1 - q_0) \geq 0,$$

where

$$p_i = p(x_i, y_i), \quad q_i = q(x_i, y_i) \quad i = 0, 1.$$

Consider the *transformation of Lewy*:

$$T(x, y) = (\xi(x, y), \eta(x, y)) = (x + p(x, y), y + q(x, y)).$$

If we set

$$\xi_i = \xi(x_i, y_i), \quad \eta_i = \eta(x_i, y_i) \quad i = 0, 1,$$

then equation (1) implies that

$$(\xi_1 - \xi_0)^2 + (\eta_1 - \eta_0)^2 \geq (x_1 - x_0)^2 + (y_1 - y_0)^2.$$

Hence $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is distance-increasing, and, in particular, T is one-one. Moreover, the Jacobian of T is

$$\begin{pmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \xi}{\partial y} \\ \frac{\partial \eta}{\partial x} & \frac{\partial \eta}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 + r & s \\ s & 1 + t \end{pmatrix},$$

with determinant

$$\begin{aligned} 1 + r + t + rt - s^2 &= 2 + r + t \quad \text{by } (*) \\ &\geq 2, \end{aligned}$$

so T is an immersion, and image T is open. But image T is also closed: For if $T(x_i, y_i) \rightarrow \alpha \in \mathbb{R}^2$, so that $\{T(x_i, y_i)\}$ is a Cauchy sequence, then $\{(x_i, y_i)\}$ is also a Cauchy sequence, since T is distance-increasing; thus $(x_i, y_i) \rightarrow \beta \in \mathbb{R}^2$, and $T(\beta) = \alpha$. So T is actually a diffeomorphism of \mathbb{R}^2 onto itself. It will be convenient to use classical ambiguous notation and denote the inverse map T^{-1} by $(\xi, \eta) \mapsto (x(\xi, \eta), y(\xi, \eta))$. Its Jacobian is

$$\begin{aligned} \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix} &= \begin{pmatrix} 1+r & s \\ s & 1+t \end{pmatrix}^{-1} \\ &= \frac{1}{2+r+t} \begin{pmatrix} 1+t & -s \\ -s & 1+r \end{pmatrix}, \end{aligned}$$

from which we can read off the partial derivatives of x and y .

Now define $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$\begin{aligned} F(\xi, \eta) &= (U(\xi, \eta), V(\xi, \eta)) \\ &= (x - p, -y + q) \quad \text{i.e.,} \\ &= (x(\xi, \eta) - p(x(\xi, \eta), y(\xi, \eta)), -y(\xi, \eta) + q(x(\xi, \eta), y(\xi, \eta))). \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial U}{\partial \xi} &= \frac{\partial x}{\partial \xi} - \frac{\partial p}{\partial x} \frac{\partial x}{\partial \xi} - \frac{\partial p}{\partial y} \frac{\partial y}{\partial \xi} \\ &= \frac{1}{2+r+t} [1+t - r(1+t) - s(-s)] \\ &= \frac{t-r}{2+r+t}. \end{aligned}$$

Similarly, we find that

$$\frac{\partial V}{\partial \eta} = \frac{t-r}{2+r+t} = \frac{\partial U}{\partial \xi}$$

and

$$\frac{\partial V}{\partial \xi} = \frac{2s}{2+r+t} = -\frac{\partial U}{\partial \eta}.$$

Thus (U, V) satisfies the Cauchy-Riemann equations, so the map $F: \mathbb{C} \rightarrow \mathbb{C}$ defined by

$$\begin{aligned} F(\xi + i\eta) &= U(\xi, \eta) + iV(\xi, \eta) \\ &= x - p + (-y + q)i \end{aligned}$$

is complex analytic, and for the complex derivative F' we have

$$\begin{aligned} (2) \quad F'(\xi + i\eta) &= \frac{\partial U}{\partial \xi} + i \frac{\partial V}{\partial \xi} \\ &= \frac{t - r + 2is}{2 + r + t}. \end{aligned}$$

Consequently,

$$\begin{aligned} |F'(\xi + i\eta)|^2 &= \frac{(t - r)^2 + 4s^2}{(2 + r + t)^2} \\ &= \frac{(t - r)^2 + 4rt - 4}{(2 + r + t)^2} \quad \text{by } (*) \\ &= \frac{(t + r)^2 - 4}{(2 + r + t)^2} = \frac{-2 + r + t}{2 + r + t}, \end{aligned}$$

which gives

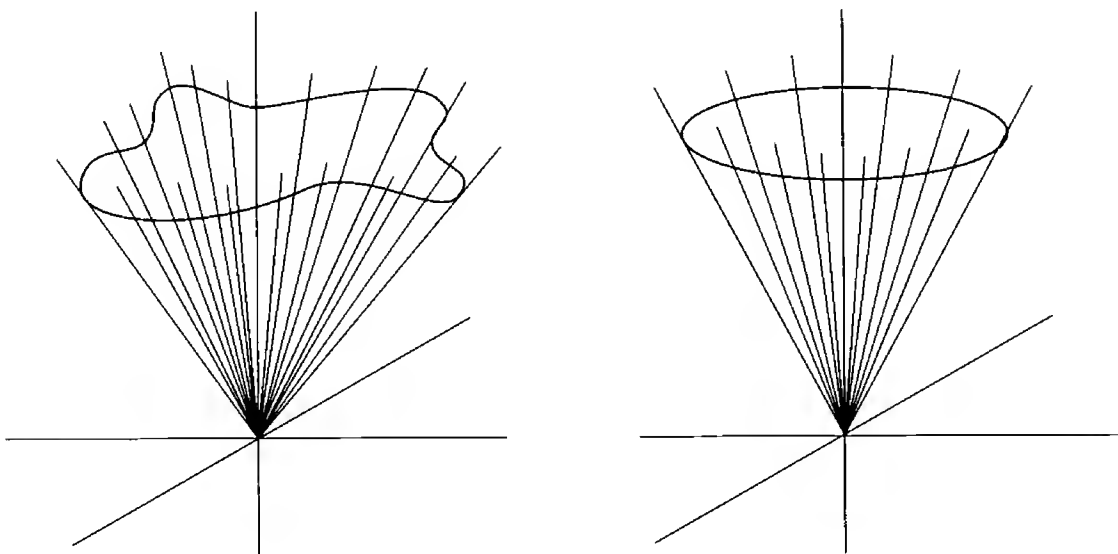
$$(3) \quad 1 - |F'(\xi + i\eta)|^2 = \frac{4}{2 + r + t} > 0.$$

Thus F' is bounded, and consequently constant, by Liouville's theorem. But equations (2) and (3) allow us to solve for r, s, t in terms of F' (here Re and Im represent the real and imaginary parts):

$$\begin{aligned} s &= \frac{2 + r + t}{2} \cdot \text{Im } F' = \frac{2 \cdot \text{Im } F'}{1 - |F'|^2} \\ \left. \begin{aligned} t - r &= \frac{4 \text{Re } F'}{1 - |F'|^2} \\ t + r &= \frac{4}{1 - |F'|^2} - 2 \end{aligned} \right\} \Rightarrow \begin{aligned} t &= \frac{1}{2} \left(\frac{4 \text{Re } F'}{1 - |F'|^2} + \frac{4}{1 - |F'|^2} - 2 \right) \\ r &= \frac{1}{2} \left(\frac{4}{1 - |F'|^2} - 2 - \frac{4 \text{Re } F'}{1 - |F'|^2} \right). \end{aligned}$$

Since F' is constant, so are r, s, t . ♦

Applying Jörgens' Theorem to our ϕ , we find that the l_{ij} are *constants*. We can assume, moreover, that $l_{12} = 0$, by means of an orthogonal transformation of \mathbb{R}^2 . Then $l_{11} = k_1$ and $l_{22} = k_2$ are the principal curvatures of the immersed surface $f(\mathbb{R}^2)$, and $k_1 k_2 = 1$. By Theorem 21, the immersion f is determined, up to an isometry of H^3 , by the pair $\{k_1, k_2\}$, with $k_1, k_2 > 0$. So in order to determine all such f , we just have to find one for each pair $\{k_1, k_2\}$ with $k_1 k_2 = 1$. For $k_1 = k_2 = 1$, all points are umbilics, and f must be a horosphere. To describe the other examples, consider the upper half-space model \mathcal{H}^3 . Our immersed surface $M \subset \mathcal{H}^3$ with constant k_1, k_2 must have isometries of \mathcal{H}^3 taking any point to any other. Now one simple case of isometries of \mathcal{H}^3 are the inversions with respect to a sphere around 0. These isometries take rays through 0 into themselves, and thus take cones through 0 into themselves. Moreover, if



we consider only right circular cones, then there are clearly isometries of \mathcal{H}^3 taking any point on a circle parallel to the (x, y) -plane to any other point on this circle, and hence there are isometries of \mathcal{H}^3 taking any point on the cone to any other point. These cones thus have constant k_1, k_2 . A simple calculation shows, in fact, that if the generators of the cone make an angle of θ with the z -axis, then the principal curvature k_1 for the principal vectors pointing along the generators is

$$k_1 = \sin \theta,$$

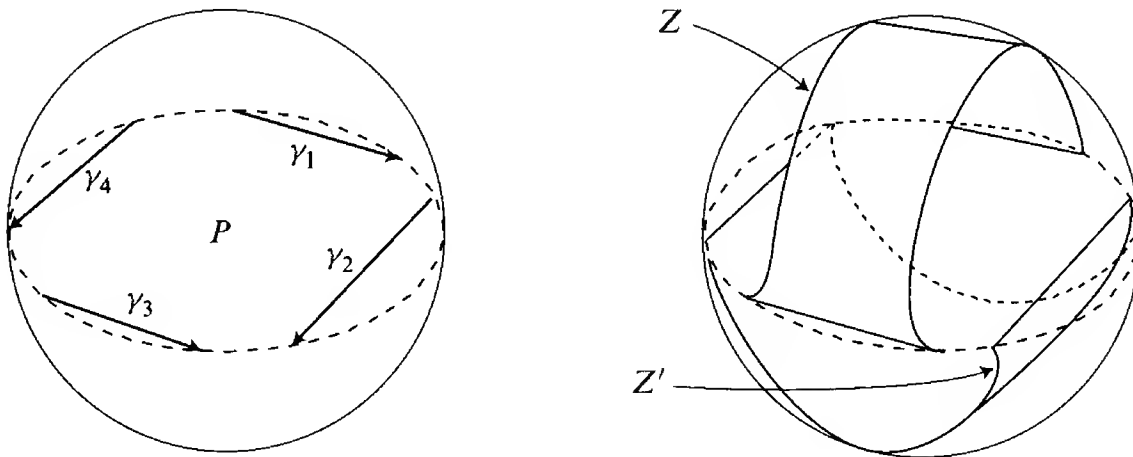
while the principal curvature k_2 for the principal vectors pointing along the circles parallel to the (x, y) -plane is

$$k_2 = \frac{1}{\sin \theta}.$$

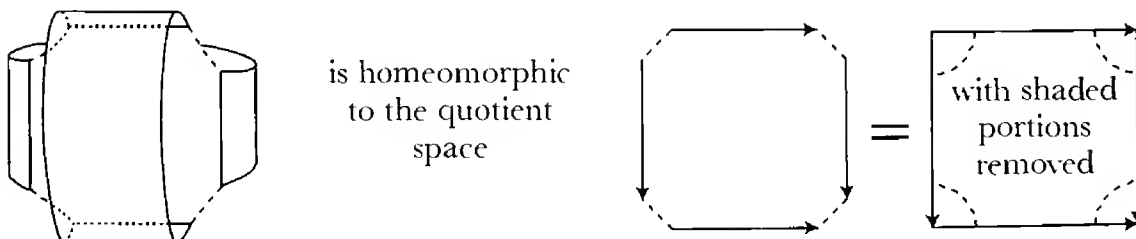
Thus $k_1 k_2 = 1$, and all pairs (k_1, k_2) are accounted for. We note, finally, that by the discussion on pages 14–15, our cone is the set of points at a fixed distance from the z -axis. We thus have

46. THEOREM (VOLKOV AND VLADIMIROVA; SASAKI). A complete surface in H^3 with constant $K_{\text{ext}} = 1$ is either a horosphere or the set of points at a fixed distance from a geodesic.

Next we consider the lower range $K_{\text{ext}} = 0 \implies K_{\text{int}} = -1$. We have already indicated that there are many complete surfaces $M \subset H^3$ with $K_{\text{ext}} = 0$, but now we will look more closely at their topological type. We know that if $B \subset \mathbb{R}^3$ is the projective model of H^3 (so B is the unit ball with a metric of constant curvature -1 whose geodesics are reparameterized straight lines of \mathbb{R}^3), then a surface $M \subset B$ has $K_{\text{int}} = -1$ if and only if M is flat, considered as a surface in \mathbb{R}^3 with the usual metric. Consider the intersection of a plane with B , and a portion P of this plane which is bounded by four non-intersecting geodesics $\gamma_1, \dots, \gamma_4$. The geodesics γ_1 and γ_3 can be joined by a cylinder Z ,



and similarly γ_2 and γ_4 can be joined by a disjoint cylinder Z' . By choosing appropriate profile curves for these cylinders we can make a smooth surface $P \cup Z \cup Z'$, and it will have $K_{\text{int}} = -1$ everywhere. The resulting surface is topologically equivalent to a torus minus a disc (or a torus minus a point).

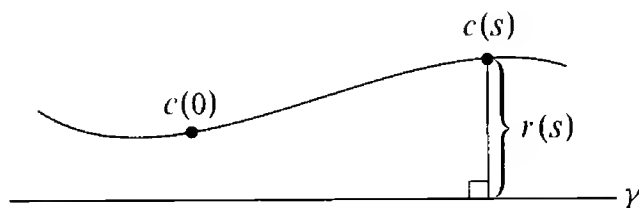


More generally, we can begin with portions P that are bounded by $2g$ non-intersecting geodesics. In this way we can obtain surfaces homeomorphic to any compact surface with a point deleted.

Notice that this construction produces only C^∞ surfaces, not analytic ones. It seems to me that all *analytic* flat surfaces in \mathbb{R}^3 , and *a posteriori* all complete analytic surfaces in B with $K_{\text{int}} = -1$, must be homeomorphic to a plane, cylinder, or Möbius strip; but I haven't tried to make a rigorous proof. If this does indeed turn out to be the case, it will be one of the rare instances where the requirements of smoothness and analyticity lead to different geometric conclusions.

We are still left with the complete surfaces in H^3 with $0 < K_{\text{ext}} < 1$. We can obtain infinitely many examples of such surfaces by looking at surfaces of revolution.

Given a geodesic γ in the hyperbolic plane, we can describe a complete arc-length parameterized curve $s \mapsto c(s)$ in the hyperbolic plane in terms of the distance $r(s)$ from $c(s)$ to γ . A curve c can be found with a given function r



provided that $|r'| \leq 1$, so that $|r(s_1) - r(s_0)| \leq s_1 - s_0$. If we rotate c around γ in hyperbolic 3-space, then the first fundamental form of our surface is (Problem 22)

$$I = \sinh^2 r(s) d\theta \otimes d\theta + ds \otimes ds,$$

and we compute that its intrinsic curvature is

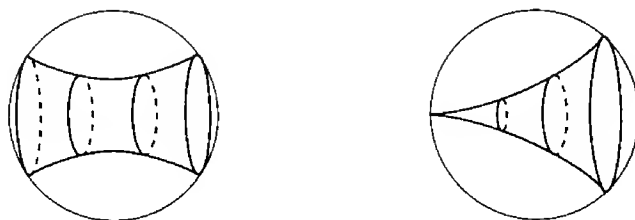
$$K_{\text{int}} = -\frac{1}{\sinh r(s)} \cdot \frac{d^2 \sinh r(s)}{ds^2}.$$

Setting $K_{\text{int}} = -c^2$, we obtain the strictly positive solution

$$\sinh r(s) = e^{cs}, \quad \text{as well as} \quad \sinh r(s) = a \cosh(cs), \quad a > 0.$$

Both solutions satisfy $|r'| < 1$ for $0 < c < 1$. Thus for each $K_{\text{ext}} = 1 - c^2$ with $0 < K_{\text{ext}} < 1$ we obtain a 1-parameter family of distinct surfaces, and

one extra one. In the model $(B^3, \langle \cdot, \cdot \rangle)$ these surfaces look like those below. I do



not know whether these are the only surfaces with $0 < K_{\text{ext}} < 1$, as in the case of surfaces with $K_{\text{ext}} = 1$, or if there are many others, as in the case $K_{\text{ext}} = 0$.

G. HYPERSURFACES OF CONSTANT CURVATURE IN HIGHER DIMENSIONS

We now want to consider hypersurfaces $M^n \subset N^{n+1}$, where $(N, \langle \cdot, \cdot \rangle)$ is a manifold of constant curvature K_0 and of dimension > 3 . We are interested in the hypersurfaces M of constant curvature; since M is no longer a surface, there is no ambiguity of meaning here—we are requiring that M , with the induced metric, have all sectional curvatures equal. After the exertions of the last section, it is a relief to find that everything is now much *easier*, and most of the results are essentially *local*. For example, we claim that there is no 3-dimensional manifold $M \subset \mathbb{R}^4$ with constant curvature -1 , not even a non-complete one. In fact, if M has principal curvatures k_1, k_2, k_3 at p , then all products $k_i k_j$ must $= -1$, which is clearly impossible, since at least two of the k_i will have the same sign. More generally,

47. THEOREM. For $n > 2$, let N^{n+1} be a manifold of constant curvature K_0 , and let $M^n \subset N^{n+1}$ be a hypersurface of constant curvature K . Then $K \geq K_0$. If $K > K_0$, then all points of M are umbilics, and if $K = K_0$, then at most one principal curvature is non-zero.

PROOF. Let k_1, \dots, k_n be the principal curvatures at p . Gauss' equation shows that

$$K - K_0 = k_i k_j, \quad i \neq j.$$

If $K = K_0$, then $k_i k_j = 0$ for all $i \neq j$, so if k_1 , say, is $\neq 0$, then $k_2, \dots, k_n = 0$. If $K - K_0 \neq 0$, then all $k_i \neq 0$, so the equation

$$k_1 k_i = k_1 k_j \quad i, j \neq 1$$

implies that $k_2 = \dots = k_n$. Similarly, $k_1 = \dots = k_{n-1}$. Since $n > 2$, this implies that $k_1 = \dots = k_n$. So all points are umbilics. Moreover, $K - K_0 = k_1 k_2 = (k_1)^2 > 0$. ♦

We will examine the case $K = K_0$ in more detail later on. But first we indicate how more general results can be obtained by considering a certain covariant tensor of order 2, the **Ricci tensor** Ric of M . The map $\text{Ric}(p): M_p \times M_p \rightarrow \mathbb{R}$ is defined by

$$\text{Ric}(p)(X_1, X_2) = \text{trace } Y \mapsto R(X_2, Y)X_1 \quad Y \in M_p.$$

In terms of the components R^i_{jkl} of R in a coordinate system x^1, \dots, x^n , the tensor Ric is given by

$$\text{Ric} = \sum_{j,k=1}^n \text{Ric}_{jk} dx^j \otimes dx^k \quad \text{for} \quad \text{Ric}_{jk} = \sum_{i=1}^n R^i_{jki};$$

thus Ric is obtained from R by contraction. If X_1, \dots, X_n is an orthonormal basis for M_p , then $\text{Ric}(X_1, X_1)$ is the trace of the matrix $(\langle R(X_1, X_i)X_1, X_j \rangle)$. Therefore

$$\begin{aligned} \text{Ric}(X_1, X_1) &= \sum_{i=1}^n \langle R(X_1, X_i)X_1, X_i \rangle \\ &= - \sum_{i=2}^n \langle R(X_i, X_1)X_1, X_i \rangle. \end{aligned}$$

So if $X \in M_p$ is any unit vector, then $-\text{Ric}(X, X)$ is the sum of the sectional curvatures determined by X and any $n-1$ orthonormal vectors orthogonal to X . (We have defined Ric so that it agrees with the classical definition $\text{Ric}_{jk} = \sum_i R^i_{jki}$; nowadays, the opposite sign is often used.) The following result is analogous to Schur's Theorem (II.7-19).

48. **THEOREM.** If M is a connected Riemannian manifold of dimension $n \geq 3$ and

$$\text{Ric}(X, Y) = \lambda \langle X, Y \rangle$$

for some function λ on M , then λ is constant.

PROOF. Bianchi's second identity (II.5-9), together with Ricci's Lemma, gives

$$(1) \quad 0 = R_{hijk;l} + R_{hikl;j} + R_{hilj;k} = 0.$$

Multiply by $\sum_{h,i,j,k} g^{hj} g^{ik}$. We have

$$\begin{aligned}
\sum g^{hj} g^{ik} R_{hijk;l} &= - \sum g^{hj} g^{ik} R_{ihjk;l} = - \sum_{h,j} g^{hj} \sum_k R^k_{hjk;l} \\
&= - \sum_{h,j} (g^{hj} \text{Ric}_{hj})_{;l} = - \sum_{h,j} (g^{hj} g_{hj} \lambda)_{;l} \\
&= -n \frac{\partial \lambda}{\partial x^l}, \\
\sum g^{hj} g^{ik} R_{hikl;j} &= \sum g^{hj} g^{ik} R_{ihlk;j} = \sum_{h,j} g^{hj} \sum_k R^k_{hik;j} \\
&= \sum_{h,j} (g^{hj} \text{Ric}_{hl})_{;j} = \sum_{h,j} (g^{hj} g_{hl} \lambda)_{;j} \\
&= \frac{\partial \lambda}{\partial x^l}, \\
\sum g^{ik} g^{hj} R_{hilj;k} &= \sum g^{ik} g^{hj} R_{jljh;k} = \sum_{i,k} g^{ik} \sum_h R^h_{lih;k} \\
&= \sum_{i,k} (g^{ik} \text{Ric}_{li})_{;k} = \sum_{i,k} (g^{ik} g_{li} \lambda)_{;k} \\
&= \frac{\partial \lambda}{\partial x^l}.
\end{aligned}$$

So (I) becomes

$$(n-2) \frac{\partial \lambda}{\partial x^l} = 0.$$

Since $n > 2$, we have $\partial \lambda / \partial x^l = 0$ for all l . ♦

A Riemannian manifold M with $\text{Ric} = -\lambda \langle \cdot, \cdot \rangle$ is called an **Einstein space**, and λ is sometimes called its **mean curvature** (not to be confused with the mean curvature H of a submanifold). If M has constant curvature K , then M is an Einstein space with mean curvature $\lambda = (n-1)K$. We note in passing that

49. THEOREM. A connected 3-dimensional Einstein space is a manifold of constant curvature.

PROOF. Choose an orthonormal basis $X_1, X_2, X_3 \in M_p$, and let $K_{ij} = K_{ji}$ be the sectional curvature of the 2-dimensional subspace of M_p spanned by X_i and X_j . Then

$$\begin{aligned}
-\text{Ric}(X_1, X_1) &= K_{12} + K_{13} \\
-\text{Ric}(X_2, X_2) &= K_{21} + K_{23} \\
-\text{Ric}(X_3, X_3) &= K_{31} + K_{32}.
\end{aligned}$$

Hence

$$-\text{Ric}(X_1, X_1) - \text{Ric}(X_2, X_2) + \text{Ric}(X_3, X_3) = 2K_{12}.$$

Since all $\text{Ric}(X_i, X_i) = -\lambda$, we have $K_{12} = \lambda/2$. ♦

Now for a manifold $(N^{n+1}, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 we consider hypersurfaces $M \subset N$ which are Einstein spaces.

50. THEOREM. For $n > 2$, let N^{n+1} be a manifold of constant curvature K_0 , and let $M^n \subset N^{n+1}$ be a hypersurface which is an Einstein space with $\text{Ric} = -\lambda \langle \cdot, \cdot \rangle$. If $\lambda > (n-1)K_0$, then all points of M are umbilics, and M is a manifold of constant curvature $K > K_0$. If $\lambda = (n-1)K_0$, then at most one principal curvature is non-zero, and M is a manifold of constant curvature K_0 .

PROOF. Let $X_1, \dots, X_n \in M_p$ be principal directions with corresponding principal curvatures k_1, \dots, k_n . Gauss' equation gives

$$(1) \quad K_{ij} = k_i k_j + K_0,$$

where K_{ij} is the sectional curvature of the subspace of M_p spanned by X_i and X_j . Then

$$\begin{aligned} \lambda &= \sum_{j \neq i} K_{ij} = \sum_{j \neq i} k_i k_j + (n-1)K_0 \\ &= \left(\sum_j k_j \right) k_i - (k_i)^2 + (n-1)K_0. \end{aligned}$$

Hence all principal curvatures k_i satisfy the equation

$$(*) \quad x^2 - \left(\sum_j k_j \right) x + [\lambda - (n-1)K_0] = 0.$$

If $\lambda = (n-1)K_0$, then every k_i is either 0 or the number $\sum_j k_j$. So there can clearly be only one $k_i \neq 0$. Then equation (1) shows that all K_{ij} equal K_0 , so that M is a manifold of constant curvature K_0 .

If $\lambda > (n-1)K_0$, then all k_i are one of the two roots α, β of (*), where

$$(2) \quad \alpha\beta = \lambda - (n-1)K_0 > 0$$

$$(3) \quad \alpha + \beta = \sum_j k_j.$$

If p of the k_i equal α , and the other $q = n - p$ of the k_i equal β , then equation (3) can be written

$$\alpha + \beta = p\alpha + q\beta \implies (p - 1)\alpha + (q - 1)\beta = 0.$$

But α and β have the same sign, by (2), so either $p - 1$ or $q - 1$ is negative, which means that either p or q is zero. Thus all k_i are equal. Then equation (1) shows that all K_{ij} equal $K_0 + (k_1)^2 > K_0$, so that M is a manifold of constant curvature $K > K_0$. ♦

Theorem 50 is not the best that can be obtained, for it is also known that if $\lambda < (n - 1)K_0$, then K_0 must be > 0 , and that in this case M must be one of a certain special class of hypersurfaces, with $\lambda = (n - 2)K_0$. For the proof of this, the reader is referred to the original paper of Fialkow [1]; see also Ryan [1].

We now consider the critical case of an immersion $f : M^n \rightarrow N^{n+1}$, where M and N have the same constant curvature K_0 , so that at most one principal curvature of $f(M)$ is non-zero at each point. If all principal curvatures are zero everywhere, so that the second fundamental form $s = 0$, then M is totally geodesic. Otherwise, we can consider the non-empty open set $U \subset M$ defined by

$$U = \{p \in M : s \neq 0 \text{ at } p\}.$$

Around any point $p \in U$, we choose an adapted orthonormal moving frame

$$X_1, \dots, X_{n-1}, X_n, X_{n+1}$$

such that X_1, \dots, X_{n-1} are principal vectors with principal curvatures 0, and X_n is a principal vector with non-zero principal curvature λ . [Our moving frame is really defined in a neighborhood of $f(p) \in N$, but for simplicity we will regard $M \subset N$ for all local arguments.] Let $\phi^\alpha, \psi_\beta^\alpha$ be the forms for this moving frame, and let θ^i, ω_j^i be the forms for X_1, \dots, X_n . Then we have

$$(1) \quad \psi_i^{n+1} = 0 \quad i = 1, \dots, n - 1$$

$$(2) \quad \psi_n^{n+1} = -\lambda\theta^n.$$

For $i = 1, \dots, n - 1$, the Codazzi-Mainardi equations give

$$0 = d\psi_i^{n+1} = \Psi_i^{n+1} - \sum_{j=1}^n \psi_j^{n+1} \wedge \omega_i^j = \Psi_i^{n+1} + \lambda\theta^n \wedge \omega_i^n.$$

Since $\Psi_i^{n+1}(X, Y) = 0$ for X, Y tangent to M , we find that $\theta^n \wedge \omega_i^n = 0$, or

$$(3) \quad \omega_i^n \text{ is a multiple of } \theta^n \text{ (on } U) \quad i = 1, \dots, n - 1.$$

From this we derive a higher dimensional analogue of Proposition 5-4.

51. PROPOSITION. The distribution Δ on $U \subset M$ defined by

$$\begin{aligned}\Delta(p) &= \{X \in M_p : s(X, Y) = 0 \text{ for all } Y \in M_p\} \\ &= \{X \in M_p : X \text{ is a principal vector with principal curvature } 0\}\end{aligned}$$

is integrable. Every integral manifold of Δ is a totally geodesic submanifold of M , and is immersed as a totally geodesic submanifold in N .

PROOF. Locally Δ is defined by $d\theta^1 = \cdots = d\theta^{n-1} = 0$. Now on U we have

$$d\theta^i = - \sum_{j=1}^n \omega_j^i \wedge \theta^j = - \sum_{j=1}^{n-1} \omega_j^i \wedge \theta^j \quad \text{by (3).}$$

So Proposition I.7-14 shows that Δ is integrable; the vector fields X_1, \dots, X_{n-1} are tangent along an integral manifold M_1 of Δ . Equation (3) says that

$$0 = \omega_i^n(X_j) = \langle \nabla'_{X_j} X_i, X_n \rangle \quad i, j \leq n-1,$$

i.e., that the second fundamental form of M_1 in M is zero. Since we also have

$$\langle \nabla'_{X_j} X_i, X_{n+1} \rangle = 0 \quad i, j \leq n-1$$

by the definition of Δ , it follows that M_1 is totally geodesic in N . ♦

Now we want to study the function λ along a geodesic γ lying in an integral manifold M_1 of Δ .

52. LEMMA. Let γ be an arclength parameterized geodesic in an integral manifold M_1 of Δ , and let $\lambda(s)$ be the value of the non-zero principal curvature at $\gamma(s)$. Then the function $\lambda(s)$ satisfies the differential equation

$$\left(\frac{1}{\lambda}\right)'' = -\frac{K_0}{\lambda}.$$

PROOF. Choose the moving frame so that γ is an integral curve of X_1 . Equations (2) and (3) imply that there are g_i with

$$(3') \quad \omega_i^n = g_i \psi_n^{n+1}.$$

The Codazzi-Mainardi equation for $i = n$ gives

$$\begin{aligned}(4) \quad d\psi_n^{n+1} &= \Psi_n^{n+1} - \sum_{j=1}^n \psi_j^{n+1} \wedge \omega_n^j \\ &= \Psi_n^{n+1}, \quad \text{by (1).}\end{aligned}$$

Thus

$$\begin{aligned}\Psi_n^{n+1} &= d\psi_n^{n+1} = -d\lambda \wedge \theta^n - \lambda d\theta^n \quad \text{by (2)} \\ &= -d\lambda \wedge \theta^n + \lambda \sum_{i=1}^n \omega_i^n \wedge \theta^i,\end{aligned}$$

and therefore

$$d\lambda \wedge \theta^n = -\lambda \sum_{i=1}^n \theta^i \wedge \omega_i^n - \Psi_n^{n+1}.$$

Applying this to (X_1, X_n) gives

$$\begin{aligned}X_1(\lambda) &= -\lambda \omega_1^n(X_n) + \lambda \omega_1^n(X_1) \\ &= -\lambda \omega_1^n(X_n) \quad \text{by (3)} \\ &= -\lambda g_1 \psi_n^{n+1}(X_n) \quad \text{by (3')} \\ &= \lambda^2 g_1 \quad \text{by (2),}\end{aligned}$$

which we can also write as

$$(*) \quad X_1 \left(\frac{1}{\lambda} \right) = -g_1.$$

Now on M we have the structural equation

$$d\omega_1^n = -\sum_{k=1}^{n-1} \omega_k^n \wedge \omega_1^k + \Omega_1^n,$$

which by (3') becomes

$$d(g_1 \psi_n^{n+1}) = -\sum_{k=1}^{n-1} g_k \psi_n^{n+1} \wedge \omega_1^k + \Omega_1^n,$$

and thus by (2) and (4)

$$dg_1 \wedge \psi_n^{n+1} + g_1 \Psi_n^{n+1} = -\lambda \left(\sum_{k=1}^{n-1} g_k \omega_1^k \right) \wedge \theta^n + \Omega_1^n.$$

Finally, we use (2) again to write our equation as

$$-\lambda dg_1 \wedge \theta^n + g_1 \Psi_n^{n+1} = -\lambda \left(\sum_{k=1}^{n-1} g_k \omega_1^k \right) \wedge \theta^n + \Omega_1^n.$$

Applying this to (X_1, X_n) we get

$$-\lambda X_1(g_1) + 0 = 0 - K_0,$$

since all $\omega_1^k(X_1) = 0$. Thus (*) yields

$$X_1 \left(X_1 \left(\frac{1}{\lambda} \right) \right) = X_1(-g_1) = -\frac{K_0}{\lambda}. \quad \blacklozenge$$

The solutions of the equation $(1/\lambda)'' = -K_0(1/\lambda)$ can be found explicitly— $1/\lambda$ is linear if $K_0 = 0$, a linear combination of \sin and \cos if $K_0 > 0$, and a linear combination of \sinh and \cosh if $K_0 < 0$. In any case, $1/\lambda$ is bounded on any bounded interval.

53. COROLLARY. If M is complete, then the integral manifolds of Δ are complete.

PROOF. We just have to show that a geodesic of an integral manifold M_1 cannot approach a boundary point of U . The argument is almost the same as that in the proof of Corollary 5-6. ♦

It is now a straightforward matter to generalize Theorem 5-9. We will make things easy for ourselves by choosing the simplest proof.

54. THEOREM. If M is a complete flat n -manifold and $f: M \rightarrow \mathbb{R}^{n+1}$ is an isometric immersion, then $f(M)$ is a generalized cylinder (it is congruent to a set of the form $\gamma \times \mathbb{R}^{n-1}$ for some curve $\gamma \subset \mathbb{R}^2$).

PROOF. We can assume M is simply connected, and thus \mathbb{R}^n . If $f(M)$ is not totally geodesic, then the set $U \subset M$ is non-empty, so by Corollary 53 some hyperplane of M is mapped isometrically onto an $(n-1)$ -dimensional plane of \mathbb{R}^{n+1} . Now apply the third proof of Theorem 5-9. ♦

We also obtain complete information for the case $K_0 > 0$.

55. THEOREM. If M^n is a complete manifold of constant curvature 1 and $f: M^n \rightarrow S^{n+1}$ is an isometric immersion, then $f(M)$ is a great n -sphere in S^{n+1} .

PROOF. We can assume M is simply connected, and thus S^n . If $f(M)$ were not totally geodesic, then the set $U \subset M$ would be non-empty, so Corollary 53 would show that there are two *disjoint* complete totally geodesic $(n-1)$ -dimensional submanifolds of $M = S^n$. This is impossible. ♦

In the case $K_0 < 0$ we would not expect such good results, since even the case $n = 2$ is so complicated. Actually, the case $n = 2$ already contains essentially all the complexity there is, for one can show that if M^n is a complete manifold

of constant curvature -1 in H^{n+1} , then the higher dimensional cohomology vanishes,

$$H^i(M) = 0 \quad \text{for } i > 1.$$

This is essentially a consequence of the analysis already provided, although technical details are required for a rigorous proof (see O'Neil [1]).

ADDENDUM 1

THE LAPLACIAN

The material of these first 3 Addenda is essentially a part of intrinsic Riemannian geometry, and might thus seem out of place in this chapter. But I felt it was appropriate to put it here since this is the first time in a long while that we have seriously considered higher dimensional Riemannian manifolds. Moreover, the next chapter will be devoted to material which is completely intrinsic in nature. Finally, some of the material covered here will be used when we return to the study of extrinsic geometry in Chapter 9.

In classical “vector analysis”, there are three operators which play a crucial role. First of all, for every smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ we have a vector field, the gradient of f , defined by

$$\text{grad } f = \left(\frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^n} \right) = \sum_{i=1}^n \frac{\partial f}{\partial x^i} \cdot \frac{\partial}{\partial x^i}.$$

On the other hand, for every vector field X on \mathbb{R}^n , with

$$X = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i},$$

we have a function, the divergence of X , defined by

$$\text{div } X = \sum_{i=1}^n \frac{\partial a_i}{\partial x^i}.$$

Finally, the Laplacian of f is the function*

$$\Delta f = \text{div}(\text{grad } f) = \sum_{i=1}^n \frac{\partial^2 f}{\partial (x^i)^2}.$$

* Classically, one introduced the operator $\nabla = \sum_i \frac{\partial}{\partial x^i} \cdot e_i$, where $e_i = \partial/\partial x^i$ is the i^{th} basis vector of \mathbb{R}^n , and wrote (formally)

$$\begin{aligned} \text{grad } f &= \nabla f \\ \text{div } X &= \langle \nabla, X \rangle = \nabla \cdot X \\ \Delta f &= \langle \nabla, \nabla f \rangle = \nabla \cdot \nabla f. \end{aligned}$$

For this reason Δ was often denoted by ∇^2 .

The operators grad, div, and Δ all have natural generalizations to an arbitrary Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$. Consider first the gradient of f . Notice that the components of grad f on \mathbb{R}^n are just the coefficients of df in the expression $df = \sum_i (\partial f / \partial x^i) dx^i$. Consequently,

$$\left\langle \text{grad } f, \sum_{i=1}^n b^i \frac{\partial}{\partial x^i} \right\rangle = \sum_{i=1}^n \frac{\partial f}{\partial x^i} b^i = df \left(\sum_{i=1}^n b^i \frac{\partial}{\partial x^i} \right).$$

We can use this equation in any Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ to define grad f as the unique vector field such that

$$(I) \quad \langle \text{grad } f, Y \rangle = df(Y) = Y(f),$$

for all vector fields Y on M . We easily see that

$$(I) \quad \text{grad}(fg) = f \cdot \text{grad } g + g \cdot \text{grad } f.$$

In terms of a coordinate system x^1, \dots, x^n on M we have

$$\text{grad } f = \sum_{i=1}^n \left(\sum_{j=1}^n g^{ij} \frac{\partial f}{\partial x^j} \cdot \frac{\partial}{\partial x^i} \right).$$

The divergence of a vector field X on M may be defined as

$$(II) \quad (\text{div } X)(p) = \text{trace } Y \mapsto \nabla_Y X \quad Y \in M_p$$

$$= \sum_{i=1}^n \langle \nabla_{Y_i} X, Y_i \rangle \quad Y_1, \dots, Y_n \in M_p \text{ orthonormal.}$$

This clearly coincides with the original definition in Euclidean space. It is easy to check that

$$(2) \quad \text{div}(fX) = X(f) + f \cdot \text{div } X = df(X) + f \cdot \text{div } X.$$

In terms of a coordinate system x^1, \dots, x^n we have

$$X = \sum_{i=1}^n a^i \frac{\partial}{\partial x^i} \implies \text{div } X = \sum_{i=1}^n a^i_{;i} = \sum_{i=1}^n \left(\frac{\partial a^i}{\partial x^i} + \sum_{j=1}^n a^j \Gamma_{ji}^i \right).$$

We can also define div ω when ω is a 1-form, for the connection ∇ on vector fields gives rise to a connection ∇ on 1-forms (Chapter II.6), and we can set

$$(III) \quad (\text{div } \omega)(p) = \sum_{i=1}^n (\nabla_{X_i} \omega)(X_i) \quad X_1, \dots, X_n \in M_p \text{ orthonormal.}$$

It is easily checked that this definition does not depend on the choice of the orthonormal basis X_1, \dots, X_n , but we can also give a completely invariant definition. We note that every bilinear map $\alpha: V \times V \rightarrow \mathbb{R}$ gives rise to a map $\alpha': V \rightarrow V^*$ by $\alpha'(v)(w) = \alpha(v, w)$. If we also have an inner product on V , then we have an isomorphism $V^* \rightarrow V$, and thus we obtain a linear map

$$V \xrightarrow{\alpha'} V^* \rightarrow V.$$

It is easily seen that the trace of this composition is the same as

$$\sum_{i=1}^n \alpha(X_i, X_i) \quad X_1, \dots, X_n \text{ an orthonormal basis for } V.$$

To apply this to the case at hand, we consider the tensor $\nabla\omega$, with

$$(\nabla\omega)(X, Y) = (\nabla_X\omega)(Y).$$

Then

$$(\operatorname{div} \omega)(p) = \text{trace of the composition } M_p \xrightarrow{(\nabla\omega)'} M_p^* \rightarrow M_p,$$

where the isomorphism $M_p^* \rightarrow M_p$ comes from the metric. The analogue of equation (2) is

$$(3) \quad \operatorname{div}(f\omega) = \langle df, \omega \rangle + f \operatorname{div} \omega,$$

where the inner product $\langle \cdot, \cdot \rangle$ on M_p^* comes from the inner product $\langle \cdot, \cdot \rangle$ on M_p in the standard way.

More generally, consider a tensor A which is covariant of order k . We define $\operatorname{div} A$ to be a covariant tensor of order $k - 1$ by the (admittedly asymmetric) formula

$$(III') \quad \operatorname{div} A(p)(Y_2, \dots, Y_k) = \sum_{i=1}^n (\nabla_{X_i} A)(X_i, Y_2, \dots, Y_k) \\ X_1, \dots, X_n \text{ orthonormal in } M_p.$$

The reader may easily work out a completely invariant definition.

In Problem I.9-13 we introduced the Divergence Theorem for n -dimensional submanifolds-with-boundary of \mathbb{R}^n . Now that we have generalized the definition of div , we would like to generalize this theorem also. An examination of the proof hinted at in that problem leads us to hope that the following alternative definition of div is valid (the symbol \sqcup is defined in Problem I.7-4).

56. LEMMA. Let M be an oriented n -dimensional Riemannian manifold, with volume element dV (which can be considered to be an n -form, since M is oriented). Then for every vector field X on M we have

$$(*) \quad d(X \lrcorner dV) = (\operatorname{div} X) \cdot dV.$$

PROOF. If $(*)$ holds for X_1 and X_2 , then it clearly holds for $X_1 + X_2$. Moreover, if $(*)$ holds for X , then

$$\begin{aligned} d(fX \lrcorner dV) &= d(f \cdot (X \lrcorner dV)) \\ &= df \wedge (X \lrcorner dV) + f \cdot d(X \lrcorner dV) \\ &= df \wedge (X \lrcorner dV) + f \cdot \operatorname{div} X \cdot dV. \end{aligned}$$

Now Problem I.7-4(f) gives

$$\begin{aligned} 0 &= X \lrcorner (df \wedge dV) = (X \lrcorner df) \wedge dV - df \wedge (X \lrcorner dV) \\ &= X(f) \cdot dV - df \wedge (X \lrcorner dV), \end{aligned}$$

so our formula becomes

$$\begin{aligned} d(fX \lrcorner dV) &= X(f) \cdot dV + f \cdot \operatorname{div} X \cdot dV \\ &= (\operatorname{div} fX) \cdot dV \quad \text{by (2).} \end{aligned}$$

Thus $(*)$ is also true for fX .

Now let X_1, \dots, X_n be a positively oriented orthonormal moving frame, with dual forms $\theta^1, \dots, \theta^n$, so that $dV = \theta^1 \wedge \dots \wedge \theta^n$. By the considerations of the previous paragraph, it suffices to prove $(*)$ when X is some X_i , and we might as well take $X = X_1$. We easily see that

$$X_1 \lrcorner dV = X_1 \lrcorner (\theta^1 \wedge \dots \wedge \theta^n) = \theta^2 \wedge \dots \wedge \theta^n.$$

So

$$\begin{aligned} d(X_1 \lrcorner dV) &= d(\theta^2 \wedge \dots \wedge \theta^n) = \sum_{j=2}^n (-1)^j \theta^2 \wedge \dots \wedge d\theta^j \wedge \dots \wedge \theta^n \\ &= - \sum_{j=2}^n (-1)^j \theta^2 \wedge \dots \wedge \left(\sum_{i=1}^n \omega_i^j \wedge \theta^i \right) \wedge \dots \wedge \theta^n \\ &= - \sum_{j=2}^n (-1)^j \theta^2 \wedge \dots \wedge (\omega_1^j \wedge \theta^1) \wedge \dots \wedge \theta^n \\ &= - \sum_{j=2}^n (-1)^j \omega_1^j \wedge \theta^1 \wedge \dots \wedge \widehat{\theta^j} \wedge \dots \wedge \theta^n. \end{aligned}$$

But

$$\omega_1^j = \sum_{k=1}^n \omega_1^j(X_k) \cdot \theta^k = \sum_{k=1}^n \langle \nabla_{X_k} X_1, X_j \rangle \theta^k,$$

so we obtain

$$\begin{aligned} d(X_1 \lrcorner dV) &= \sum_{j=2}^n \langle \nabla_{X_j} X_1, X_j \rangle \theta^1 \wedge \cdots \wedge \theta^n \\ &= (\operatorname{div} X_1) dV. \quad \spadesuit \end{aligned}$$

As an easy corollary we now obtain

57. THEOREM (THE DIVERGENCE THEOREM). Let M be a compact oriented n -dimensional Riemannian manifold-with-boundary, with outward pointing unit normal v on ∂M . Denote the volume element of M by dV_n , and that of ∂M by dV_{n-1} . Let X be a vector field on M . Then

$$\int_M \operatorname{div} X \, dV_n = \int_{\partial M} \langle X, v \rangle \, dV_{n-1}.$$

PROOF. This follows from Stokes' Theorem, and the easily verified fact that $X \lrcorner dV_n$ equals $\langle X, v \rangle dV_{n-1}$ on ∂M . \spadesuit

58. COROLLARY (GREEN'S THEOREM). If M is a compact oriented n -dimensional Riemannian manifold without boundary, and X is any vector field on M , then

$$\int_M \operatorname{div} X \, dV_n = 0.$$

Notice that even when M is not orientable, equation (*) in Lemma 56 can be used to define $\operatorname{div} X$, for both sides of the equation change sign when the orientation is reversed, so locally the formula defines $\operatorname{div} X$ unambiguously.

We now define the Laplacian Δf of f on M by

$$(IV) \quad \Delta f = \operatorname{div}(\operatorname{grad} f).$$

For a coordinate system x^1, \dots, x^n on M we have, with the notation of Chapter II.5,

$$\begin{aligned}\Delta f &= \sum_{i=1}^n \left(\sum_{j=1}^n g^{ij} \frac{\partial f}{\partial x^j} \right)_{;i} = \sum_{i=1}^n \left(\sum_{j=1}^n g^{ij} f_{;j} \right)_{;i} \\ &= \sum_{i,j=1}^n g^{ij} f_{;ji} = \sum_{i,j=1}^n g^{ij} f_{;ij} \\ &= \sum_{i,j=1}^n g^{ij} \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \sum_{k=1}^n \frac{\partial f}{\partial x^k} \Gamma_{ij}^k \right).\end{aligned}$$

If x^1, \dots, x^n is a normal coordinate system at p , so that $\Gamma_{ij}^k(p) = 0$, and $g_{ij}(p) = \delta_{ij}$, then

$$\Delta f(p) = \sum_{i=1}^n \frac{\partial^2 f}{\partial (x^i)^2}(p).$$

We can also state a more precise result along these lines. Suppose that X_1, \dots, X_n are vector fields which are orthonormal at p . Then

$$\begin{aligned}(\text{IV}') \quad \Delta f(p) &= \text{div}(\text{grad } f)(p) \\ &= \text{trace } X \mapsto \nabla_X \text{grad } f \quad X \in M_p \\ &= \sum_{i=1}^n \langle \nabla_{X_i(p)} \text{grad } f, X_i(p) \rangle \\ &= \sum_{i=1}^n X_i(p) \langle \text{grad } f, X_i \rangle - \sum_{i=1}^n \langle (\text{grad } f)(p), \nabla_{X_i(p)} X_i \rangle \\ &= \sum_{i=1}^n X_i(p) (df(X_i)) - \sum_{i=1}^n \langle (\text{grad } f)(p), \nabla_{X_i(p)} X_i \rangle \\ &= \sum_{i=1}^n (X_i X_i f)(p) - \sum_{i=1}^n \langle (\text{grad } f)(p), \nabla_{X_i(p)} X_i \rangle.\end{aligned}$$

So we have

$$(\text{IV}'') \quad \Delta f(p) = \sum_{i=1}^n (X_i X_i f)(p) \quad \text{for} \quad \begin{cases} X_1, \dots, X_n \text{ orthonormal at } p \\ \nabla_{X_i} X_i = 0 \text{ at } p. \end{cases}$$

We ought to mention that the Laplacian Δf on a surface was first introduced by Beltrami, so Δ is often called the Laplace-Beltrami operator. For reasons that

will appear in the next Addendum, the Laplacian is often defined as the negative of the Laplacian as defined here. There is no general agreement on the proper sign, so whenever a lecturer states that her Laplacian is the usual one (or the negative of the usual one), one half of the audience (or the other half) raises their eyebrows and murmurs disgruntledly “hmmph, so she calls *that* the usual Laplacian!”

A simple calculation [using normal coordinates, or equation (IV''), to make things even easier] shows that

$$(4) \quad \Delta(fg) = f \cdot \Delta g + g \cdot \Delta f + 2\langle \text{grad } f, \text{grad } g \rangle.$$

We will use this formula to derive a result of importance later on.

59. PROPOSITION. Let M be a compact oriented n -dimensional Riemannian manifold-with-boundary, with outward pointing unit normal v on ∂M . Then

$$\int_M [f \Delta f + \langle \text{grad } f, \text{grad } f \rangle] dV_n = \int_{\partial M} \langle f \text{grad } f, v \rangle dV_{n-1}.$$

In particular, if $f = 0$ on ∂M [and, *a posteriori* if $\partial M = \emptyset$], then

$$\int_M f \Delta f dV_n = - \int_M \langle \text{grad } f, \text{grad } f \rangle dV_n.$$

PROOF. The Divergence Theorem (Theorem 57) gives

$$\int_M \Delta(f^2) dV_n = \int_M \text{div}(\text{grad } f^2) dV_n = \int_{\partial M} \langle \text{grad } f^2, v \rangle dV_{n-1}.$$

Then equations (1) and (4) give the result. ♦

As a corollary we have

60. LEMMA (BOCHNER'S LEMMA). Let M be a compact connected Riemannian manifold (without boundary). If $f: M \rightarrow \mathbb{R}$ has $\Delta f \geq 0$ everywhere, then f is a constant function (and $\Delta f = 0$).

PROOF. We can assume that M is orientable, by taking the orientable 2-fold covering space of M if necessary. First of all, Corollary 58 gives

$$\int_M \Delta f dV = \int_M \text{div}(\text{grad } f) dV = 0.$$

Since $\Delta f \geq 0$ on M , this already implies that $\Delta f = 0$ on M . Now (the second part of) Proposition 59 gives

$$0 = \int_M f \Delta f \, dV = - \int_M \langle \text{grad } f, \text{grad } f \rangle \, dV.$$

So we must have $\text{grad } f = 0 \implies df = 0 \implies f$ is constant. \blacklozenge

An alternative proof of Lemma 60 is given in Addendum 2 to Chapter 10.

A couple of explicit calculations of the Laplacian will be used at various times. Our first calculation is most easily carried out in a coordinate system. Consider a 1-form

$$\omega = \sum_i a_i \, dx^i,$$

and the vector field

$$X = \sum_i a^i \frac{\partial}{\partial x^i}, \quad a^i = \sum_j g^{ij} a_j.$$

This vector field X is described intrinsically by the equation

$$\langle X, Y \rangle = \omega(Y) \quad \text{for all vector fields } Y,$$

so that, in particular,

$$\langle X, X \rangle = \omega(X) = \sum_i a^i a_i.$$

Now

$$\begin{aligned} \Delta \left(\sum_i a^i a_i \right) &= \sum_{j,k} g^{jk} \left(\sum_i a^i a_i \right)_{;jk} \\ &= \sum_{j,k} g^{jk} \sum_i (a^i_{;j} a_i + a^i a_{i;j})_{;k} \\ &= \sum_{j,k} g^{jk} \sum_i (a^i_{;jk} a_i + a^i_{;j} a_{i;k} + a^i_{;k} a_{i;j} + a^i a_{i;jk}). \end{aligned}$$

Since

$$\begin{aligned} \sum_{i,j,k} g^{jk} a^i_{;jk} a_i &= \sum_{i,j,k} \sum_{l,m} g^{jk} g^{il} a_{l;jk} g_{im} a^m \\ &= \sum_{j,k,m} g^{jk} a_{m;jk} a^m = \sum_{i,j,k} g^{jk} a_{i;jk} a^i, \end{aligned}$$

and

$$\sum_{i,j,k} g^{jk} a^i_{,j} a_{i;k} = \sum_{i,j,k} \sum_l g^{jk} g^{il} a_{l;j} a_{i;k},$$

we have, finally,

$$(5) \quad \Delta \left(\sum_i a^i a_i \right) = 2 \left(\sum_{i,j,k} g^{jk} a^i a_{i;jk} + \sum_{i,j,k,l} g^{jk} g^{il} a_{i;k} a_{l;j} \right).$$

Our second calculation is easily carried out in a coordinate-free way. Consider an immersion $f: M^n \rightarrow \mathbb{R}^m$, so that M has a Riemannian metric $I_f = f^*\langle \cdot, \cdot \rangle$. We will compute Δf with respect to this metric (the fact that f is \mathbb{R}^m -valued causes no difficulty, for we can compute the Laplacian of each of its component functions—for simplicity we suppress the various components and simply use formula (IV'') for \mathbb{R}^m -valued f). It will make things conceptually easier to think of M^n as a subset of \mathbb{R}^m , so that f is the inclusion map. Let X_1, \dots, X_n be vector fields on M satisfying the conditions of (IV''). We first want to figure out what the \mathbb{R}^m -valued function $X_i(f)$ is. Now

$$\begin{aligned} X_i(f) &= df(X_i) = \text{“the vector part of” } f_*(X_i) && \text{by Problem I.4-3} \\ &= X_i && \text{(when } X_i \text{ is considered as a point of } \mathbb{R}^m\text{).} \end{aligned}$$

Therefore,

$$\begin{aligned} X_i(X_i f)(p) &= \nabla'_{X_i} X_i(p) \\ &= \nabla_{X_i} X_i(p) + s(X_i(p), X_i(p)) \\ &= s(X_i(p), X_i(p)) && \text{by our conditions on } X_1, \dots, X_n. \end{aligned}$$

Thus we see that

$$(6) \quad \Delta f(p) = n \cdot \eta(p),$$

where η is the mean curvature normal. Notice, in particular, that if M has $\eta = 0$, then $\Delta f = 0$. Lemma 60 then implies that M cannot be compact (for $n \geq 1$), which reproves Corollary 31. In the particular case of a hypersurface, we have

$$(7) \quad \Delta f = nH \cdot \nu,$$

where ν is the unit normal field.

Notice the correspondence between equation (7) and equation (II') on pg. III.109, which can be written in the simple form

$$\Delta_{\langle \cdot, \cdot \rangle} f = \mathcal{N},$$

where $\Delta_{\langle \cdot, \cdot \rangle}$ indicates the Laplacian with respect to the metric $\langle \cdot, \cdot \rangle$ on M . Since the Laplacian is such a natural operator on a Riemannian manifold, it is not surprising to find $\Delta_{\langle \cdot, \cdot \rangle} f$ related to \mathcal{N} . (Note also that $\Delta_{\langle \cdot, \cdot \rangle}$ involves g_{ij} and Christoffel symbols Γ_{ij}^k , and thus third derivatives of f , just like \mathcal{N}). As a matter of fact, this equation was originally used as the *definition* of \mathcal{N} (it is clearly a special linear affine invariant!).

The Laplacian can be generalized in two very important ways. One such generalization is treated in the next Addendum. A different generalization, important in Chapter 9, is suggested by the next to last line of equation (IV'), which can be written

$$\Delta f(p) = \sum_{i=1}^n X_i(p)(df(X_i)) - \sum_{i=1}^n df(\nabla_{X_i(p)} X_i)$$

X_1, \dots, X_n orthonormal at p .

Now Corollary II.6-5 says that the covariant derivative ∇df is given by

$$(\nabla_{X_p} df)(Y_p) = X_p(df(Y)) - df(\nabla_{X_p} Y)$$

for any vector fields X, Y extending X_p, Y_p . Thus we can write

$$\Delta f(p) = \sum_{i=1}^n (\nabla_{X_i} df)(X_i) \quad X_1, \dots, X_n \in M_p \text{ orthonormal.}$$

Thus, using (III) we can just as well define Δf by

$$\Delta f = \operatorname{div}(df).$$

[Naturally, one could, with some work, demonstrate the equation $\operatorname{div}(\operatorname{grad} f) = \operatorname{div}(df)$ directly from the completely invariant definitions.]

The nice thing about this new definition of Δf is that it can be generalized immediately. Consider a vector bundle $\varpi: E \rightarrow M$, where M has a metric $\langle \cdot, \cdot \rangle$, and E has some connection D . If ξ is any section of E , then $D_{X_p} \xi \in \varpi^{-1}(p)$ for $X_p \in M_p$. We can therefore think of $D\xi$ as a section of the bundle $\operatorname{Hom}(TM, E)$ whose fibre at p is $\operatorname{Hom}(M_p, \varpi^{-1}(p))$. Now the connection ∇ on M determined by $\langle \cdot, \cdot \rangle$, together with the connection D on E , determines

a connection $\tilde{\nabla}$ on $\text{Hom}(TM, E)$. This is defined as on page 37, except that the situation is even simpler. As in that case, we easily see that for any vector fields X, Y and any section ψ on $\text{Hom}(TM, E)$, we have

$$(8) \quad (\tilde{\nabla}_{X_p} \psi)(Y_p) = D_{X_p}(\psi(Y)) - \psi(\nabla_{X_p} Y).$$

[If $E = M \times \mathbb{R}$, so that the sections of E are functions $f: M \rightarrow \mathbb{R}$, and we define $D_X f$ to be $df(X)$, then $\tilde{\nabla}$ will just be the connection ∇ on 1-forms.] Naturally $\tilde{\nabla}\psi$ will denote the section of $\text{Hom}(TM \times TM, E)$ with

$$(\tilde{\nabla}\psi)(X, Y) = (\tilde{\nabla}_X \psi)(Y).$$

For a section ξ of E we can now define

$$(V) \quad \Delta\xi(p) = \sum_{i=1}^n (\tilde{\nabla}_{X_i} D\xi)(X_i) \quad X_1, \dots, X_n \in M_p \text{ orthonormal.}$$

(A completely invariant definition is easily formulated, as before.) If we let X_1, \dots, X_n be vector fields which are orthonormal at p , then

$$\begin{aligned} (V') \quad \Delta\xi(p) &= \sum_{i=1}^n (\tilde{\nabla}_{X_i(p)} D\xi)(X_i(p)) \\ &= \sum_{i=1}^n D_{X_i(p)}(D\xi(X_i)) - \sum_{i=1}^n D\xi(\nabla_{X_i(p)} X_i) \quad \text{by (8)} \\ &= \sum_{i=1}^n D_{X_i(p)} D_{X_i} \xi - \sum_{i=1}^n D\xi(\nabla_{X_i(p)} X_i). \end{aligned}$$

So we have, in complete analogy with equation (IV''),

$$(V'') \quad \Delta\xi(p) = \sum_{i=1}^n (D_{X_i} D_{X_i} \xi)(p) \quad \text{for} \quad \begin{cases} X_1, \dots, X_n \text{ orthonormal at } p \\ \nabla_{X_i} X_i = 0 \text{ at } p. \end{cases}$$

ADDENDUM 2

THE $*$ OPERATOR AND THE
LAPLACIAN ON FORMS; HODGE'S THEOREM

Let V be an oriented n -dimensional vector space with an inner product $\langle \cdot, \cdot \rangle$. The $*$ operator, from alternating k -linear functions $\Omega^k(V)$ to $\Omega^{n-k}(V)$, is usually defined as follows. Let v_1, \dots, v_n be a positively oriented orthonormal basis of V , and let ϕ_1, \dots, ϕ_n be the dual basis. Then

$$*(\phi_{i_1} \wedge \cdots \wedge \phi_{i_k}) = \pm \phi_{j_1} \wedge \cdots \wedge \phi_{j_{n-k}},$$

where i_1, \dots, i_k are k distinct numbers from $1, \dots, n$, and j_1, \dots, j_{n-k} are the other $n - k$ numbers of this set, arranged in some order; we use the $+$ sign if $v_{i_1}, \dots, v_{i_k}, v_{j_1}, \dots, v_{j_{n-k}}$ is positively oriented, and the $-$ sign otherwise. We also set $*1 = \pm \phi_1 \wedge \cdots \wedge \phi_n$, where $1 \in \Omega^0(V) = \mathbb{R}$, and $*(\phi_1 \wedge \cdots \wedge \phi_n) = \pm 1$. It is easy to see, first of all, that this definition is consistent, for a fixed basis v_1, \dots, v_n , and then that the definition is also independent of the orthonormal basis. An invariant definition can be given as follows. We always have a map

$$\Omega^k(V) \times \Omega^{n-k}(V) \xrightarrow{\wedge} \Omega^n(V).$$

An orientation and inner product on V gives us an isomorphism $\Omega^n(V) \xrightarrow{\approx} \mathbb{R}$, so we have a bilinear map

$$\{ \cdot, \cdot \}: \Omega^k(V) \times \Omega^{n-k}(V) \rightarrow \mathbb{R}.$$

Then we can define

$$A: \Omega^k(V) \rightarrow (\Omega^{n-k}(V))^*$$

by

$$A(\omega)(\eta) = \{\omega, \eta\} \quad \omega \in \Omega^k(V), \quad \eta \in \Omega^{n-k}(V).$$

Now the inner product on V also gives us an isomorphism $V \rightarrow V^*$ from which we derive an isomorphism $(\Omega^{n-k}(V))^* \rightarrow \Omega^{n-k}(V)$. One easily checks that the composition

$$\Omega^k(V) \xrightarrow{A} (\Omega^{n-k}(V))^* \rightarrow \Omega^{n-k}(V)$$

is precisely $*$. Straightforward calculations show that

$$(1) \quad ** = * \circ *: \Omega^k(V) \rightarrow \Omega^k(V) \quad \text{is } (-1)^{k(n-k)} \text{ times the identity.}$$

In Chapter I.9 we mentioned that the inner product on V gives inner products on all vector spaces $\Omega^k(V)$, although we did not describe most of these inner products explicitly. The inner product on $\Omega^1(V) = V^*$ can be described by the condition that the dual basis ϕ^1, \dots, ϕ^n is orthonormal if and only if v_1, \dots, v_n is orthonormal in V . Using the inner product $\langle \cdot, \cdot \rangle$ thus defined on $\Omega^1(V)$, we can describe the inner product on $\Omega^k(V)$ as the unique one with

$$(2) \quad \langle \phi_1 \wedge \dots \wedge \phi_k, \psi_1 \wedge \dots \wedge \psi_k \rangle = \det(\langle \phi_i, \psi_j \rangle)$$

for $\phi_i, \psi_j \in V^*$. In particular, if ϕ_1, \dots, ϕ_n is orthonormal in V^* , then

$$\begin{aligned} & \langle \phi_{i_1} \wedge \dots \wedge \phi_{i_k}, \phi_1 \wedge \dots \wedge \phi_k \rangle \\ &= \det \begin{pmatrix} \delta_{i_1 1} & \dots & \delta_{i_1 k} \\ \vdots & & \vdots \\ \delta_{i_k 1} & \dots & \delta_{i_k k} \end{pmatrix} \\ &= \begin{cases} 0 & \text{if } \{i_1, \dots, i_k\} \neq \{1, \dots, k\} \\ \text{sgn } \pi & \text{if } i_\alpha = \pi(\alpha) \text{ for some permutation } \pi \text{ of } \{1, \dots, k\}. \end{cases} \end{aligned}$$

Since the naming of the indices was purely arbitrary, we have, just as well,

$$(3) \quad \langle \phi_{i_1} \wedge \dots \wedge \phi_{i_k}, \phi_{j_1} \wedge \dots \wedge \phi_{j_k} \rangle = \begin{cases} 0 & \text{if } \{i_1, \dots, i_k\} \neq \{j_1, \dots, j_k\} \\ \text{sgn } \pi & \text{if } j_\alpha = \pi(i_\alpha). \end{cases}$$

So we can also describe the inner product on $\Omega^k(V)$ as the one which makes the $\phi_{i_1} \wedge \dots \wedge \phi_{i_k}$ ($i_1 < \dots < i_k$) an orthonormal basis, for any orthonormal basis ϕ_1, \dots, ϕ_n of V^* .

Now note that

$$\begin{aligned} \phi_{i_1} \wedge \dots \wedge \phi_{i_k} \wedge *(\phi_1 \wedge \dots \wedge \phi_k) &= \phi_{i_1} \wedge \dots \wedge \phi_{i_k} \wedge \pm \phi_{k+1} \wedge \dots \wedge \phi_n \\ &= \begin{cases} 0 & \text{if } \{i_1, \dots, i_k\} \neq \{1, \dots, k\} \\ (\text{sgn } \pi) \cdot *1 & \text{if } i_\alpha = \pi(\alpha). \end{cases} \end{aligned}$$

Again, since the naming of the indices was arbitrary, we have, just as well,

$$(4) \quad \phi_{i_1} \wedge \dots \wedge \phi_{i_k} \wedge *(\phi_{j_1} \wedge \dots \wedge \phi_{j_k}) = \begin{cases} 0 & \text{if } \{i_1, \dots, i_k\} \neq \{j_1, \dots, j_k\} \\ (\text{sgn } \pi) \cdot *1 & \text{if } j_\alpha = \pi(i_\alpha). \end{cases}$$

Comparing (3) and (4), we see that for $\omega, \eta \in \Omega^k(V)$ we have

$$(5) \quad \langle \omega, \eta \rangle = *(\omega \wedge *\eta) = *(\eta \wedge *\omega).$$

Now everything that we have done can be extended to k -forms on an oriented Riemannian n -manifold $(M, \langle \cdot, \cdot \rangle)$. We have an operator $*$ taking k -forms to $(n - k)$ -forms, and $** = (-1)^{k(n-k)}$ on k -forms. It is easy to check (using the dual forms to an orthonormal moving frame, for example) that $*$ takes C^∞ forms to C^∞ forms. Note that the volume element dV on M is just $*1$ for the constant function (0-form) 1.

We also have, for two k -forms, ω and η , a function $\langle \omega, \eta \rangle$ on M . We would like a formula for $\langle \omega, \eta \rangle$ when we have coordinate expressions

$$(a) \quad \omega = \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k}$$

$$(b) \quad \eta = \sum_{j_1 < \dots < j_k} b_{j_1 \dots j_k} dx^{j_1} \wedge \dots \wedge dx^{j_k}.$$

For this, and later, purposes, it will be convenient to express a form in terms of tensor products of the dx^i , instead of wedge products. Recall (Theorem I.7-2(3)) that

$$\begin{aligned} dx^{i_1} \wedge \dots \wedge dx^{i_k} &= \frac{(1 + \dots + 1)!}{1! \dots 1!} \text{Alt}(dx^{i_1} \otimes \dots \otimes dx^{i_k}) \\ &= \sum_{\sigma \in S_k} \text{sgn } \sigma \, dx^{\sigma(i_1)} \otimes \dots \otimes dx^{\sigma(i_k)}. \end{aligned}$$

This shows that the expression (a) can also be written

$$\omega = \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k},$$

where the new $a_{i_1 \dots i_k}$ are skew-symmetric in the indices i_1, \dots, i_k and agree with the old $a_{i_1 \dots i_k}$ when $i_1 < \dots < i_k$. Now let g_{ij} be the components of $\langle \cdot, \cdot \rangle$ in our coordinate system, so that g^{ij} are the components of $\langle \cdot, \cdot \rangle$ on the dual space. With any tensor, covariant of order k ,

$$A = \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k},$$

we can associate the tensor, contravariant of order k ,

$$\tilde{A} = \sum_{j_1, \dots, j_k} a^{j_1 \dots j_k} \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_k}},$$

where

$$(6) \quad a^{j_1 \dots j_k} = \sum_{i_1, \dots, i_k} g^{i_1 j_1} \dots g^{i_k j_k} a_{i_1 \dots i_k}.$$

In the special case where

$$A = \sum_i a_i dx^i,$$

it is clear how \tilde{A} is described invariantly: if we think of $\tilde{A}(p)$ as a linear function on M_p^* , then

$$\tilde{A}(p)(\phi) = A(p)(S(\phi)),$$

where $S: M_p^* \rightarrow M_p$ is the isomorphism given by the metric. In general,

$$\tilde{A}(p)(\phi_1, \dots, \phi_k) = A(p)(S(\phi_1), \dots, S(\phi_k)).$$

Notice that if the $a_{i_1 \dots i_k}$ are skew-symmetric in the indices, then so are the $a^{j_1 \dots j_k}$. So if ω is given by (a), then $\tilde{\omega}$ is also given by

$$\tilde{\omega} = \sum_{j_1 < \dots < j_k} a^{j_1 \dots j_k} \frac{\partial}{\partial x^{j_1}} \wedge \dots \wedge \frac{\partial}{\partial x^{j_k}}$$

[note, however, that the $a^{j_1 \dots j_k}$ are computed from (6), in which $a_{i_1 \dots i_k}$ is defined, by skew-symmetry, for all i_1, \dots, i_k]. We now claim that for ω, η given by (a) and (b), we have

$$(7) \quad \begin{aligned} \langle \omega, \eta \rangle &= \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} b^{i_1 \dots i_k} = \sum_{i_1 < \dots < i_k} a^{i_1 \dots i_k} b_{i_1 \dots i_k} \\ &= \frac{1}{k!} \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} b^{i_1 \dots i_k} = \frac{1}{k!} \sum_{i_1, \dots, i_k} a^{i_1 \dots i_k} b_{i_1 \dots i_k}. \end{aligned}$$

To prove this, we note that the last two expressions can be defined invariantly as contractions (traces) of $\omega \otimes \tilde{\eta}$ or $\tilde{\omega} \otimes \eta$. So it suffices to check that (7) holds at a point p where dx^1, \dots, dx^n are orthonormal. In this case $g^{ij} = \delta^{ij}$ at p , so $a^{i_1 \dots i_k} = a_{i_1 \dots i_k}$ at p . The desired result then follows immediately from equation (3).

On the oriented Riemannian n -manifold M we can also do something else. Since we have the map d , which raises the degree of a form, we can define a map δ , which lowers the degree of a form, by

$$\delta = (-1)^{n(k+1)+1} * d * \quad \text{from } k\text{-forms to } (k-1)\text{-forms.}$$

We clearly have $\delta^2 = 0$, and $\delta = 0$ on functions (0-forms). Note that on k -forms we have

$$\begin{aligned}
 (8) \quad * \delta &= (-1)^{n(k+1)+1} (**) d* \\
 &= (-1)^{n(k+1)+1} \cdot (-1)^{(n-k+1)(k-1)} d* \\
 &\quad \text{by (1) [since } d* \text{ of a } k\text{-form is an } n-k+1 \text{ form]} \\
 &= (-1)^k d*,
 \end{aligned}$$

and similarly

$$(9) \quad \delta* = (-1)^{k+1} *d.$$

Notice that δ can really be defined even when M is not orientable, for its definition is local, and changing the orientation of M reverses the sign of $*$, so leaves δ unchanged. We now define an operator Δ from k -forms to k -forms by

$$\Delta = \delta d + d\delta.$$

The reader may check that on 0-forms this Δ is the negative of the one in the previous Addendum. [N. B. The connection ∇ on M gives rise in a natural way to a connection ∇ on the bundle of k -forms on M , so the final definition of the previous Addendum also gives us a Laplacian on k -forms. But that Laplacian is *not* related to the one defined here.] Simple computations, using (8) and (9) for the last equation, give

$$(10) \quad d\Delta = \Delta d, \quad \delta\Delta = \Delta\delta, \quad *\Delta = \Delta*.$$

On a compact oriented manifold M we can define the inner product (ω, η) of two k -forms ω, η by

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle dV = \int_M \omega \wedge *\eta \quad \text{by (5).}$$

This inner product $(\ , \)$ is clearly symmetric and positive definite. Now if ω is a $(k-1)$ -form and η is a k -form, then

$$\begin{aligned}
 d(\omega \wedge *\eta) &= d\omega \wedge *\eta + (-1)^{k-1} \omega \wedge d*\eta \\
 &= d\omega \wedge *\eta - \omega \wedge *\delta\eta \quad \text{by (8).}
 \end{aligned}$$

So Stokes' Theorem gives

$$0 = \int_M d(\omega \wedge *\eta) = \int_M d\omega \wedge *\eta - \int_M \omega \wedge *\delta\eta,$$

or

$$(11) \quad (d\omega, \eta) = (\omega, \delta\eta).$$

Thus δ is the “adjoint” of d for the inner product $(\ , \)$, and this property characterizes $\delta\eta$, since $(\ , \)$ is positive definite. From this we easily see that Δ is self-adjoint with respect to the inner product $(\ , \)$ on k -forms,

$$(12) \quad (\Delta\omega, \eta) = (\omega, \Delta\eta).$$

In Euclidean space, a function f with $\Delta f = 0$ is called harmonic. In an oriented Riemannian manifold $(M, \langle \ , \ \rangle)$ we call a k -form ω **harmonic** if $\Delta\omega = 0$. When M is compact, we can write

$$(\Delta\omega, \omega) = ([d\delta + \delta d]\omega, \omega) = (\delta\omega, \delta\omega) + (d\omega, d\omega),$$

which shows that

$$(13) \quad \Delta\omega = 0 \implies d\omega = 0 \text{ and } \delta\omega = 0, \quad M \text{ compact}$$

(the converse is trivial). If ω and η are k -forms, and $\Delta\omega = 0$, then equation (12) gives

$$(\Delta\eta, \omega) = (\eta, \Delta\omega) = 0.$$

So the vector space of all harmonic k -forms (the kernel of Δ) is orthogonal to the image of Δ . The fundamental result on harmonic forms states that these two orthogonal subspaces of the k -forms span the whole vector space of k -forms:

THE HODGE DECOMPOSITION THEOREM. If M is a compact oriented Riemannian n -manifold, then for each k with $0 \leq k \leq n$, the vector space H^k of harmonic k -forms is finite dimensional, and the vector space $E^k(M)$ of all k -forms on M can be written as an orthogonal direct sum decomposition

$$E^k(M) = \Delta(E^k(M)) \oplus H^k(M).$$

For a proof of this result, which is completely analytic in nature, the reader is referred to Warner [1]; the proof given there is elementary and completely self-contained. We will merely indicate the consequences of the theorem for the

de Rham cohomology. The orthogonal decomposition of $E^k(M)$ gives two projection maps

$$\begin{array}{ccc} & & H^k(M) \\ & \nearrow H^k & \\ E^k(M) & & \\ & \searrow h^k & \\ & & \Delta E^k(M). \end{array}$$

For any $\alpha \in E^k(M)$, the form $h^k(\alpha) = \alpha - H^k(\alpha)$ is uniquely $\Delta\omega$ for some ω . Set

$$G(\alpha) = \text{the unique } \omega \text{ with } \Delta\omega = \alpha - H^k(\alpha),$$

so that

$$G = [\Delta|_{\Delta(E^k(M))}]^{-1} \circ h^k.$$

Now consider any linear map $T: E^k(M) \rightarrow E^l(M)$ with $T\Delta = \Delta T$ [e.g., $T = d, \delta, \Delta$]. We easily see that

$$T(H^k) \subset H^l, \quad T(\Delta(E^k(M))) \subset \Delta(E^l(M)).$$

So

$$T \circ h^k = h^l \circ T, \quad T \circ [\Delta|_{\Delta(E^k(M))}] = [\Delta|_{\Delta(E^l(M))}] \circ T.$$

From this we see that $GT = TG$. In particular, G commutes with d .

Now let ω be any k -form. Then we have

$$\begin{aligned} \alpha &= \Delta G\alpha + H^k(\alpha) \\ &= d\delta G\alpha + \delta dG\alpha + H^k(\alpha) \\ &= d\delta G\alpha + \delta Gd\alpha + H^k(\alpha). \end{aligned}$$

So if $d\alpha = 0$, then

$$\alpha = d\delta G\alpha + H^k(\alpha).$$

Thus $H^k(\alpha)$ is a harmonic k -form in the same de Rham cohomology class as α . On the other hand, suppose α_1 and α_2 are two harmonic k -forms in the same de Rham cohomology class, so that

$$\alpha_1 - \alpha_2 = d\beta$$

for some β . Then

$$\begin{aligned}(d\beta, d\beta) &= (d\beta, \alpha_1 - \alpha_2) = (\beta, \delta\alpha_1 - \delta\alpha_2) && \text{by (11)} \\ &= 0 && \text{by (13).}\end{aligned}$$

So $d\beta = 0$, or $\alpha_1 = \alpha_2$. Thus there is a *unique* harmonic form in each de Rham cohomology class. In other words, the k -dimensional de Rham cohomology vector space is isomorphic to the vector space $H^k(M)$ of harmonic k -forms.

We will give a simple application of this result in a moment. First we would like to observe that both d and δ can be defined in terms of the connection ∇ of M . For d this is easy.

61. PROPOSITION. If ω is a k -form on a Riemannian manifold, then

$$d\omega = (-1)^k(k+1) \cdot \text{Alt } \nabla\omega.$$

PROOF. Let x^1, \dots, x^n be a normal coordinate system at p , and let

$$\begin{aligned}\omega &= \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k} \\ &= \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k}, \quad \text{as on page 141.}\end{aligned}$$

Then (pg. II.231)

$$\nabla\omega = \sum_{i_1, \dots, i_k} \sum_h a_{i_1 \dots i_k; h} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes dx^h.$$

So

$$\begin{aligned}(k+1)! \text{Alt } \nabla\omega &= \sum_{i_1, \dots, i_k} \sum_h a_{i_1 \dots i_k; h} (k+1)! \text{Alt}(dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes dx^h) \\ &= \sum_{i_1, \dots, i_k} \sum_h a_{i_1 \dots i_k; h} dx^{i_1} \wedge \dots \wedge dx^{i_k} \wedge dx^h \\ &= k! \sum_{i_1 < \dots < i_k} \sum_h a_{i_1 \dots i_k; h} dx^{i_1} \wedge \dots \wedge dx^{i_k} \wedge dx^h.\end{aligned}$$

by skew-symmetry of the $a_{i_1 \dots i_k}$. So

$$(*) \quad (-1)^k (k+1) \text{Alt } \nabla \omega = \sum_{i_1 < \dots < i_k} \sum_h a_{i_1 \dots i_k; h} dx^h \wedge dx^{i_1} \wedge \dots \wedge dx^{i_k}.$$

But at p we have (Proposition II.5-1)

$$a_{i_1 \dots i_k; h}(p) = \frac{\partial}{\partial x^h} a_{i_1 \dots i_k}(p).$$

So

$$\begin{aligned} & (-1)^k (k+1) \text{Alt } \nabla \omega(p) \\ &= \sum_{i_1 < \dots < i_k} \sum_h \frac{\partial}{\partial x^h} a_{i_1 \dots i_k}(p) dx^h \wedge dx^{i_1} \wedge \dots \wedge dx^{i_k}(p) \\ &= d\omega(p). \quad \blacklozenge \end{aligned}$$

Naturally, the use of a normal coordinate system at p was merely a simplifying device; in an arbitrary coordinate system we would obtain the same result with a little more calculation—the Christoffel symbols in $(*)$ all cancel out after we write all $dx^h \wedge dx^{i_1} \wedge \dots \wedge dx^{i_k}$ in terms of increasing sequences of indices. The formula

$$d\omega = \sum_{i_1 < \dots < i_k} \sum_h a_{i_1 \dots i_k; h} dx^h \wedge dx^{i_1} \wedge \dots \wedge dx^{i_k},$$

which follows from $(*)$ and the final result of the theorem, can be rewritten as follows:

$$d\omega = \sum_{j_1 < \dots < j_{k+1}} b_{j_1 \dots j_{k+1}} dx^{j_1} \wedge \dots \wedge dx^{j_{k+1}},$$

for

$$b_{j_1 \dots j_{k+1}} = \sum_{\mu=1}^{k+1} (-1)^{\mu+1} a_{j_1 \dots \widehat{j_\mu} \dots j_{k+1}; j_\mu}.$$

Notice that if the a 's are skew-symmetric, then the b 's will be also.

Now suppose that we have a $(k+1)$ -form

$$\eta = \sum_{j_1 < \dots < j_{k+1}} c_{j_1 \dots j_{k+1}} dx^{j_1} \wedge \dots \wedge dx^{j_{k+1}}.$$

Then equation (7) gives

$$\begin{aligned}
 (14) \quad \langle d\omega, \eta \rangle &= \frac{1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} b_{j_1 \dots j_{k+1}} c^{j_1 \dots j_{k+1}} \\
 &= \frac{1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} \sum_{\mu=1}^{k+1} (-1)^{\mu+1} a_{j_1 \dots \widehat{j}_\mu \dots j_{k+1}; j_\mu} c^{j_1 \dots j_{k+1}} \\
 &= \frac{1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} \sum_{\mu=1}^{k+1} (-1)^{\mu+1} \cdot -a_{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} c^{j_1 \dots j_{k+1}; j_\mu} \\
 &\quad + \frac{1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} \sum_{\mu=1}^{k+1} (-1)^{\mu+1} (a_{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} c^{j_1 \dots j_{k+1}})_{; j_\mu} \\
 &= \Sigma_1 + \Sigma_2, \text{ say.}
 \end{aligned}$$

Now it is easy to see that

$$\begin{aligned}
 &\sum_{j_1, \dots, j_{k+1}} a_{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} c^{j_1 \dots j_{k+1}; j_\mu} \\
 &= \sum_{j_1, \dots, j_{k+1}} a^{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} \sum_{\rho} g^{j_\mu \rho} c_{j_1 \dots j_{k+1}; \rho} \\
 &= \sum_{j_1, \dots, j_{k+1}} (-1)^{\mu-1} a^{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} \sum_{\rho} g^{j_\mu \rho} c_{j_\mu j_1 \dots \widehat{j}_\mu \dots j_{k+1}; \rho}.
 \end{aligned}$$

Hence

$$\begin{aligned}
 \Sigma_1 &= \frac{-1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} \sum_{\mu=1}^{k+1} a^{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} \sum_{\rho} g^{j_\mu \rho} c_{j_\mu j_1 \dots \widehat{j}_\mu \dots j_{k+1}; \rho} \\
 &= \frac{-1}{(k+1)!} \sum_{j_1, \dots, j_{k+1}} \sum_{\mu=1}^{k+1} a^{j_1 \dots \widehat{j}_\mu \dots j_{k+1}} \gamma_{j_1 \dots \widehat{j}_\mu \dots j_{k+1}}, \text{ say} \\
 &= -\frac{k+1}{(k+1)!} \sum_{l_1, \dots, l_k} a^{l_1 \dots l_k} \gamma_{l_1 \dots l_k} \\
 &= -\frac{1}{k!} \sum_{l_1, \dots, l_k} a^{l_1 \dots l_k} \gamma_{l_1 \dots l_k}.
 \end{aligned}$$

Now the γ 's are simply the components of the tensor $\text{div } \eta$, defined by (III') on page 130. So

$$(15) \quad \Sigma_1 = -\langle \omega, \text{div } \eta \rangle.$$

On the other hand, we obtain $k + 1$ well-defined vector fields X_μ with

$$X_\mu = \sum_{\rho=1}^n \left(\sum_{j_1, \dots, j_{k+1}} a_{j_1 \dots \widehat{j_\mu} \dots j_{k+1}} c^{j_1 \dots j_{\mu-1} \rho j_{\mu+1} \dots j_{k+1}} \right) \frac{\partial}{\partial x^\rho}.$$

Then

$$(16) \quad \Sigma_2 = \operatorname{div} \left(\frac{1}{(k+1)!} \sum_{\mu=1}^{k+1} X_\mu \right) = \operatorname{div} Y, \text{ say.}$$

Combining (14), (15), (16), we have

$$(*) \quad \langle d\omega, \eta \rangle = -\langle \omega, \operatorname{div} \eta \rangle + \operatorname{div} Y.$$

From this we conclude

62. PROPOSITION. If η is a $(k+1)$ -form on an oriented Riemannian manifold M , then

$$\delta\eta = -\operatorname{div} \eta.$$

PROOF. First suppose that M is compact. Equation $(*)$ gives, for any k -form ω ,

$$\begin{aligned} \langle d\omega, \eta \rangle &= \int_M \langle d\omega, \eta \rangle dV = \int_M \langle \omega, -\operatorname{div} \eta \rangle dV + \int_M \operatorname{div} Y dV \\ &= \int_M \langle \omega, -\operatorname{div} \eta \rangle dV + 0 \quad \text{by Corollary 58} \\ &= \langle \omega, -\operatorname{div} \eta \rangle. \end{aligned}$$

Since $\delta\eta$ is the unique form with $\langle d\omega, \eta \rangle = \langle \omega, \delta\eta \rangle$ for all k -forms ω , it follows that $\delta\eta = -\operatorname{div} \eta$.

If M is not compact, we can still conclude, from Theorem 57, that $\langle d\omega, \eta \rangle = \langle \omega, -\operatorname{div} \eta \rangle$ for all k -forms ω with compact support. This is still sufficient to imply that $\delta\eta = -\operatorname{div} \eta$. ♦

Now consider a 1-form $\omega = \sum_i a_i dx^i$. Propositions 61 and 62 say that

$$\begin{aligned} d\omega = 0 &\iff a_{i;j} = a_{j;i} \\ \delta\omega = 0 &\iff 0 = \sum_{i,j} g^{ij} a_{i;j} = \sum_i a^i_{;i}. \end{aligned}$$

Suppose that $d\omega = \delta\omega = 0$, and consider the expression (5) on page 136 for $\Delta(\sum_i a^i a_i)$. For the first term in parentheses we have

$$\begin{aligned}
\sum_{i,j,k} g^{jk} a^i a_{i;jk} &= \sum_{i,j,k} g^{jk} a^i a_{j;ik} && \text{since } d\omega = 0 \\
&= \sum_{i,j,k} g^{jk} a^i \left(a_{j;ki} + \sum_l a_l R^l_{jik} \right) && \text{by Ricci's identity} \\
&= \sum_{i,j,k} a^i (g^{jk} a_{j,k})_{;i} + \sum_{i,j,k,l} g^{jk} a^i a_l R^l_{jik} \\
&= 0 + \sum_{i,j,k,l} g^{jk} a^i a_l R^l_{jik} && \text{since } \delta\omega = 0 \\
&= \sum_{i,j,k,l,\mu} g^{jk} a^i g_{l\mu} a^\mu R^l_{jik} \\
&= \sum_{i,j,k,\mu} a^i a^\mu g^{jk} R_{\mu jik} \\
&= - \sum_{i,j,k,\mu} a^i a^\mu g^{jk} R_{j\mu ik} \\
&= - \sum_{i,k,\mu} a^i a^\mu R^k_{\mu ik} \\
&= - \sum_{i,\mu} \text{Ric}_{\mu i} a^i a^\mu.
\end{aligned}$$

Thus we obtain

$$(*) \quad \Delta\left(\sum_i a^i a_i\right) = 2\left[- \sum_{i,j} \text{Ric}_{ij} a^i a^j + \sum_{i,j,k,l} g^{jk} g^{il} a_{i;k} a_{l;j}\right].$$

63. THEOREM (BOCHNER). Let M be a compact oriented Riemannian manifold with $-\text{Ric}(X, X) > 0$ for all $X \neq 0$. (This holds, in particular, if all sectional curvatures of M are > 0 .) Then the 1-dimensional de Rham cohomology of M is zero.

PROOF. Let ω be any 1-form on M with $\Delta\omega = 0$. Then also $d\omega = \delta\omega = 0$, by (13). Then $(*)$ shows that $\Delta(\sum_i a^i a_i) \geq 0$, since the second sum on the right is clearly ≥ 0 . Recall (page 135) that $\sum_i a^i a_i$ is a well-defined function f on M . So Lemma 60 implies that $\Delta(\sum_i a^i a_i) = 0$. By the hypothesis on Ric ,

this implies that $\omega = 0$. In other words, 0 is the only harmonic 1-form. Since the vector space of all harmonic 1-forms is isomorphic to the 1-dimensional de Rham cohomology of M , the theorem follows. ♦

In the next Chapter we will prove that a compact Riemannian manifold M satisfying $-\text{Ric}(X, X) > 0$ for all $X \neq 0$ actually has a finite fundamental group $\pi_1(M)$. Then the result of Theorem 63 follows by algebraic topology. [First we use the Hurewicz theorem to conclude that the first homology group $H_1(M; \mathbb{Z})$ is finite; then the universal coefficient theorem implies that $H^1(M; \mathbb{R}) = \text{Hom}(H_1(M; \mathbb{Z}), \mathbb{R}) = 0$]. However, Theorem 63 has many generalizations, proved using similar techniques, that have never been strengthened in the same way.

ADDENDUM 3

WHEN ARE TWO
RIEMANNIAN MANIFOLDS ISOMETRIC?

Suppose we are given two Riemannian manifolds M, \bar{M} of the same dimension n . We would like a way of finding out whether they are locally isometric. In other words, we ask if there is a point $p \in M$, a point $\bar{p} \in \bar{M}$, and an isometry $\alpha: U \rightarrow \bar{U}$ of a neighborhood U of p onto a neighborhood \bar{U} of \bar{p} . We have a slightly different problem if we are already given p and \bar{p} , and merely seek U and \bar{U} . Admittedly, both of these problems are a little strange, for we are not very likely to be given two explicit Riemannian metrics just out of the clear blue sky; specific metrics which actually come up in practice are so special, and the requirements of isometry so stringent, that there is usually no difficulty seeing whether they are isometric. As a matter of fact, I know of no instance where the (complicated) general methods which we will develop are actually used. But it is nevertheless quite significant that we can now settle the question of isometry in the category of Riemannian manifolds, for this shows that any intrinsic invariant of a Riemannian manifold can be defined in terms of the various invariants (like the curvature tensor) which we have already discovered.

The theory is rather special, and quite pleasant, in the 2-dimensional case. First some preliminaries. For two functions f, g on a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, we introduce the classical notation*

$$\Delta_1(f, g) = \langle \text{grad } f, \text{grad } g \rangle, \quad \Delta_1 f = \Delta_1(f, f).$$

It is clear that if $\alpha: M \rightarrow \bar{M}$ is an isometry, and $f, g: \bar{M} \rightarrow \mathbb{R}$, then

$$(1) \quad \bar{\Delta}_1(f \circ \alpha, g \circ \alpha) = \Delta_1(f, g),$$

where $\bar{\Delta}_1$ is formed with respect to the metric on \bar{M} . For a metric

$$\langle \cdot, \cdot \rangle = E du \otimes du + F[du \otimes dv + dv \otimes du] + G dv \otimes dv$$

on a 2-dimensional manifold, we easily compute that

$$\Delta_1(f, g) = \frac{1}{EG - F^2} \left[E \frac{\partial f}{\partial v} \frac{\partial g}{\partial v} - F \left(\frac{\partial f}{\partial v} \frac{\partial g}{\partial u} + \frac{\partial f}{\partial u} \frac{\partial g}{\partial v} \right) + G \frac{\partial f}{\partial u} \frac{\partial g}{\partial u} \right].$$

* In classical differential geometry books, the Laplacian Δf was denoted by $\Delta_2 f$ (and, worst of all, $\Delta_1 f$ was sometimes written as Δf), but we will stick with Δf for the Laplacian.

In particular, we have

$$\Delta_1 u = \frac{G}{EG - F^2}, \quad \Delta_1(u, v) = \frac{-F}{EG - F^2}, \quad \Delta_1 v = \frac{E}{EG - F^2}.$$

This gives

$$\frac{1}{EG - F^2} = \Delta_1 u \cdot \Delta_1 v - \Delta_1(u, v)^2 = \Theta^2(u, v), \quad \text{say,}$$

and thus

$$(2) \quad E = \frac{\Delta_1 v}{\Theta^2(u, v)}, \quad F = \frac{-\Delta_1(u, v)}{\Theta^2(u, v)}, \quad G = \frac{\Delta_1 u}{\Theta^2(u, v)}.$$

This equation shows that the metric $\langle \cdot, \cdot \rangle$ is determined once we know $\Delta_1(u)$, $\Delta_1(u, v)$, and $\Delta_1(v)$ for any coordinate system (u, v) . We can formalize the contents of this equation as follows.

64. LEMMA. Let $\alpha: M \rightarrow \bar{M}$ be a diffeomorphism of 2-dimensional Riemannian manifolds, and for each coordinate system (\bar{u}, \bar{v}) on \bar{M} , define $(u, v) = (\bar{u}, \bar{v}) \circ \alpha$ on M . If α is an isometry, then

$$(*) \quad \Delta_1 u = (\bar{\Delta}_1 \bar{u}) \circ \alpha, \quad \Delta_1(u, v) = \bar{\Delta}_1(\bar{u}, \bar{v}) \circ \alpha, \quad \Delta_1 v = (\bar{\Delta}_1 \bar{v}) \circ \alpha.$$

Conversely, if these equations hold for some collection of coordinate systems (\bar{u}, \bar{v}) whose domains cover \bar{M} , then α is an isometry.

PROOF. Since $u = \bar{u} \circ \alpha$ and $v = \bar{v} \circ \alpha$, the first part of the theorem follows immediately from equation (1). To prove the converse, let the metrics on M and \bar{M} be

$$\langle \cdot, \cdot \rangle = E du \otimes du + \cdots \quad \text{and} \quad \langle \cdot, \cdot \rangle^- = \bar{E} d\bar{u} \otimes d\bar{u} + \cdots.$$

Since

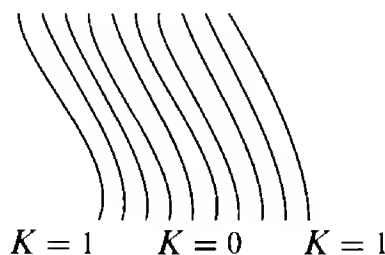
$$du = d(\bar{u} \circ \alpha) = \alpha^*(d\bar{u}), \quad dv = \alpha^*(d\bar{v}),$$

we have

$$\alpha^*(\bar{E} d\bar{u} \otimes d\bar{u} + \cdots) = (\bar{E} \circ \alpha) du \otimes du + \cdots.$$

But the hypothesis $(*)$, together with equation (2), gives $\bar{E} \circ \alpha = E$, etc. ♦

Now consider two metrics $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle^-$ on two 2-dimensional Riemannian manifolds M and \bar{M} . We want to know if there is locally an isometry $\alpha: M \rightarrow \bar{M}$. It might happen that the curvature K of $\langle \cdot, \cdot \rangle$ is constant. Then α exists if and only if the curvature \bar{K} of $\langle \cdot, \cdot \rangle^-$ is the same constant; if this is the case, then there is a 2-parameter family of isometries α . Suppose instead that the curvature K of $\langle \cdot, \cdot \rangle$ is not constant. We consider a region where the sets $K = \text{constant}$ give a foliation, and we will try to decide whether the isometry exists



in this region. To be sure, the sets $K = \text{constant}$ might look much worse; for example $K = 0$ might be a single point, or a whole set with interior, etc., etc. But in general, the more complicated the decomposition we obtain, the *easier* it will be to handle the problem, for then the sets $\bar{K} = \text{constant}$ must look just as complicated. At any rate, we will, over and over again, restrict our attention to the “general” case, and not worry about the exceptional situations. When the sets $K = \text{constant}$ give a foliation, then the sets $\bar{K} = \text{constant}$ must also, if the required isometry α is to exist. Moreover, the isometry α must take the set $K = c$ onto the set $\bar{K} = c$. However this still leaves a lot of leeway, and does not yet determine α . We now consider the function $\Delta_1 K$. This function might not give us any new information at all, for $\Delta_1 K$ might be a constant on each of the sets $K = \text{constant}$. We will first consider the case where $\Delta_1 K$ is not constant on these sets. In fact, we want to assume that $\Delta_1 K$ varies monotonically on each set $K = \text{constant}$. Then it will “generally” be the case that $(K, \Delta_1 K): M \rightarrow \mathbb{R}^2$ is a local coordinate system for M . This is the situation which we will actually consider. If the isometry α is to exist, then $(\bar{K}, \bar{\Delta}_1 \bar{K}): \bar{M} \rightarrow \mathbb{R}^2$ must also be a local coordinate system for \bar{M} . Suppose this also occurs. Then clearly the isometry α must, in fact, be the composition

$$\alpha = (\bar{K}, \bar{\Delta}_1 \bar{K})^{-1} \circ (K, \Delta_1 K).$$

defined in some open set $U \subset M$. Now the question arises: how do we know whether this α is actually an isometry? There is an easy answer to this question:

Lemma 64 tells us that α is an isometry if and only if

$$\begin{aligned}\Delta_1 K &= \bar{\Delta}_1 \bar{K} \circ \alpha \\ \Delta_1(K, \Delta_1 K) &= \bar{\Delta}_1(\bar{K}, \bar{\Delta}_1 \bar{K}) \circ \alpha \\ \Delta_1(\Delta_1 K) &= \bar{\Delta}_1(\bar{\Delta}_1 \bar{K}) \circ \alpha.\end{aligned}$$

Moreover, the first of these equations is automatic, by the definition of α .

Now consider the opposite extreme, where $\Delta_1 K$ is a function of K . If α exists, then $\bar{\Delta}_1 \bar{K}$ must be the same function of \bar{K} ,

$$(a) \quad \Delta_1 K = f \circ K, \quad \bar{\Delta}_1 \bar{K} = f \circ \bar{K}.$$

We look at the Laplacians ΔK and $\bar{\Delta} \bar{K}$. If $(K, \Delta K)$ is a local coordinate system, then $(\bar{K}, \bar{\Delta} \bar{K})$ must be also, and α must be

$$\alpha = (\bar{K}, \bar{\Delta} \bar{K})^{-1} \circ (K, \Delta K).$$

This α is an isometry if and only if

$$\Delta_1(K, \Delta K) = \bar{\Delta}_1(\bar{K}, \bar{\Delta} \bar{K}) \circ \alpha, \quad \Delta_1(\Delta K) = \bar{\Delta}_1(\bar{\Delta} \bar{K}) \circ \alpha;$$

the extra condition $\Delta_1 K = \bar{\Delta}_1 \bar{K} \circ \alpha$ follows from (a) and the definition of α . This still leaves us with the case where ΔK also fails to be independent of K in the worst possible way, so that in addition to (a) we have

$$(b) \quad \Delta K = g \circ K, \quad \bar{\Delta} \bar{K} = g \circ \bar{K}.$$

Then it turns out (Problem 24) that the surfaces are isometric, and there is a 1-parameter family of isometries between them.

For higher dimensional manifolds the treatment will be more systematic, but correspondingly less concrete. We already know (Corollary II.7-13) that the metric in a normal coordinate system determined by an orthonormal frame X_{1p}, \dots, X_{np} is completely determined by knowing $\langle R(X_i, X_j)X_k, X_l \rangle$, where X_1, \dots, X_n is the moving frame adapted to X_{1p}, \dots, X_{np} . This result gives us a criterion for determining when a neighborhood of $p \in M$ is isometric to a neighborhood of $\bar{p} \in \bar{M}$, but it cannot be regarded as a reasonable solution of our problem, for we may not be able to compute the geodesics, or the parallel translations along these geodesics. All we can compute is the *equations* for the geodesics and for parallel translations—usually we will not be able to solve these equations explicitly. What we want is a criterion involving only quantities directly computable in a coordinate system—like curvature, covariant derivatives of tensors which have already been computed, etc.

Recall the map $\Phi: \mathbb{R} \times M_p \rightarrow M$ (pg. II.270) defined by

$$\Phi(t, X_p) = \exp(tX_p).$$

From the discussion on pp. II.270–278 [c.f. especially Corollary 9 and Theorem 12] we see that the metric in the normal coordinate system determined by X_{1p}, \dots, X_{np} is completely determined once the functions $\mathbf{R}^i_{jkl} \circ \Phi$ are known. Now suppose that the metric is *analytic*. Then its form in normal coordinates is known once we know

$$\frac{\partial(\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t}(0, X_p), \quad \frac{\partial^2(\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t^2}(0, X_p), \quad \dots \quad \text{all } X_p \in M_p.$$

Now

$$(1) \quad \frac{\partial(\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t}(t, X_p) = \lim_{h \rightarrow 0} \frac{\mathbf{R}^i_{jkl}(\Phi(t+h, X_p)) - \mathbf{R}^i_{jkl}(\Phi(t, X_p))}{h}.$$

Let \mathcal{R} be the tensor

$$\mathcal{R}(X, Y, Z, W) = \langle R(X, Y)Z, W \rangle,$$

so that (c.f. pg. II.277)

$$\mathbf{R}^i_{jkl} = \mathcal{R}(X_k, X_l, X_j, X_i).$$

Since $\Phi(t, X_p) = \exp(tX_p)$, and since the X_i are defined by parallel translating the X_{ip} along geodesics, equation (1) can be written

$$\begin{aligned} & \frac{\partial(\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t}(t, X_p) \\ &= (\nabla_{X(\Phi(t, X_p))} \mathcal{R})(X_k(\Phi(t, X_p)), X_l(\Phi(t, X_p)), X_j(\Phi(t, X_p)), X_i(\Phi(t, X_p))) \\ &= (\nabla \mathcal{R})(X_k(\Phi(t, X_p)), X_l(\Phi(t, X_p)), X_j(\Phi(t, X_p)), X_i(\Phi(t, X_p)), X(\Phi(t, X_p))), \end{aligned}$$

where X is also defined by parallel translating X_p along geodesics. In general, we have

$$\begin{aligned} & \frac{\partial^d(\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t^d}(t, X_p) \\ &= (\overbrace{\nabla \cdots \nabla}^d \mathcal{R})(X_k(\Phi(t, X_p)), \dots, X_i(\Phi(t, X_p)), \overbrace{X(\Phi(t, X_p)), \dots, X(\Phi(t, X_p))}^d). \end{aligned}$$

In particular,

$$\frac{\partial^d (\mathbf{R}^i_{jkl} \circ \Phi)}{\partial t^d} (0, X_p) = (\nabla^d \mathcal{R})(X_{pk}, X_{pl}, X_{pj}, X_{pi}, \overbrace{X_p, \dots, X_p}^d).$$

Thus, in the analytic case, the metric in normal coordinates around p is determined completely by knowing all $\nabla^d \mathcal{R}$ at p .

Thus we have a criterion for deciding when some neighborhood of a given point $p \in M$ can be taken isometrically onto a neighborhood of a point $\bar{p} \in \bar{M}$. This criterion works only for analytic metrics, but its real defect is the fact that we have to compute infinitely many quantities $(\nabla^d \mathcal{R})(p)$. Now we will explain how one can decide whether some open set in M is isometric to some open set of \bar{M} , without being given points p, \bar{p} in advance, without assuming the metric is analytic, and by computing only finitely many covariant derivatives $\nabla^d \mathcal{R}$. True, we will have to compute the $\nabla^d \mathcal{R}$ on all of M , not just at one point p , but in practice the only way to compute the $(\nabla^d \mathcal{R})(p)$ is to compute the $\nabla^d \mathcal{R}$ in a whole coordinate system anyway.

First we consider a general problem having nothing to do with metrics. Suppose we are given two manifolds M^N and \bar{M}^N of the same dimension N . Let $\omega^1, \dots, \omega^N$ be N everywhere linearly independent 1-forms on (a subset of) M , and let $\bar{\omega}^1, \dots, \bar{\omega}^N$ be similar 1-forms on \bar{M} . We will find a way of deciding when there is locally a diffeomorphism $\alpha: M \rightarrow \bar{M}$ such that $\omega^i = \alpha^* \bar{\omega}^i$ for $i = 1, \dots, N$. First of all, let us write

$$\begin{aligned} d\omega^i &= \sum_{j < k} C_{jk}^i \omega^j \wedge \omega^k \\ d\bar{\omega}^i &= \sum_{j < k} \bar{C}_{jk}^i \bar{\omega}^j \wedge \bar{\omega}^k \end{aligned}$$

for certain functions C_{jk}^i and \bar{C}_{jk}^i . If α exists, then we will have $C_{jk}^i = \bar{C}_{jk}^i \circ \alpha$. Now suppose that among the functions C_{jk}^i there are N which form a coordinate system (u_1, \dots, u_N) on M . Then if α exists, the corresponding N functions \bar{C}_{jk}^i must form a coordinate system $(\bar{u}_1, \dots, \bar{u}_N)$ on \bar{M} . In this case, the diffeomorphism α must be

$$(1) \quad \alpha = (\bar{u}_1, \dots, \bar{u}_N)^{-1} \circ (u_1, \dots, u_N).$$

For this α we certainly have

$$(2) \quad C_{jk}^i = \bar{C}_{jk}^i \circ \alpha$$

when C_{jk}^i is one of the u 's, and we may add the extra condition that equation (2) hold in all cases. [In the "general" case, the other C_{jk}^i will be functions of the u 's,

$$C_{jk}^i = f_{jk}^i \circ (u_1, \dots, u_N),$$

so we are demanding that the other \bar{C}_{jk}^i be the same functions of the \bar{u} 's.] We still have to decide when α given by (1) is the required diffeomorphism. For this we write

$$\begin{aligned} dC_{jk}^i &= \sum_l C_{jk,l}^i \omega^l \\ d\bar{C}_{jk}^i &= \sum_l \bar{C}_{jk,l}^i \bar{\omega}^l \end{aligned}$$

for certain functions $C_{jk,l}^i$ and $\bar{C}_{jk,l}^i$. If α has the desired properties, then we must also have

$$(3) \quad C_{jk,l}^i = \bar{C}_{jk,l}^i \circ \alpha.$$

Conversely, suppose equation (3) holds. Then

$$\begin{aligned} (4) \quad \sum_l C_{jk,l}^i \cdot (\omega^l - \alpha^* \bar{\omega}^l) &= \sum_l C_{jk,l}^i \omega^l - \sum_l (\bar{C}_{jk,l}^i \circ \alpha) \cdot \alpha^* \bar{\omega}^l \\ &= \sum_l C_{jk,l}^i \omega^l - \alpha^* \left(\sum_l \bar{C}_{jk,l}^i \cdot \bar{\omega}^l \right) \\ &= dC_{jk}^i - \alpha^*(d\bar{C}_{jk}^i) \\ &= dC_{jk}^i - d(\bar{C}_{jk}^i \circ \alpha) \\ &= 0 \quad \text{by (2).} \end{aligned}$$

This is a set of N^3 equations in N unknowns. It can be written in terms of the $N^3 \times N$ matrix $(C_{jk,l}^i)$ in which l denotes the column, and $_{jk}^i$ denotes the row. This matrix contains the $N \times N$ submatrix $(u_{i,l})$, which is non-singular, since (u_1, \dots, u_N) is a coordinate system. So the matrix $(C_{jk,l}^i)$ has rank N . This means that the only solution of our equations is the zero solution. Thus, $\omega^l = \alpha^* \bar{\omega}^l$ for $l = 1, \dots, N$.

Suppose, on the contrary, that we can choose only $N_1 < N$ functions C_{jk}^i which are independent (meaning that for any coordinate system x^1, \dots, x^N , the $N_1 \times N$ matrix $(\partial C_{jk}^i / \partial x^l)$ has rank N_1 ; or equivalently, that the $N_1 \times N$ matrix $(C_{jk,l}^i)$ has rank N_1). We now look at the functions $C_{jk,l}^i$. Among these we may

be able to choose N_2 functions $C_{jk,l}^i$ which together with the N_1 functions C_{jk}^i are independent. If $N_1 + N_2 < N$, then we look at the functions $C_{jk,lm}^i$ defined by

$$dC_{jk,l}^i = \sum_m C_{jk,lm}^i \omega^m.$$

Among these we may be able to pick N_3 which can be added to the $N_1 + N_2$ functions already obtained. Suppose that, after continuing in this way, we eventually obtain $N_1 + \dots + N_\mu = N$ independent functions. Then we can determine what α must be; moreover, we can decide whether this α really works by seeing if α satisfies

$$C_{jk,l_1 \dots l_{\mu+1}}^i = \bar{C}_{jk,l_1 \dots l_{\mu+1}}^i \circ \alpha.$$

On the other hand, it may happen that we never obtain N independent functions. In the general case this will happen because at some stage, the functions

$$C_{jk,l_1 \dots l_{\mu+1}}^i$$

are all functions of the previously chosen functions. [Notice that once this happens at stage μ , it will happen at all later stages. So in general, the integers N_1, N_2, \dots which we picked in the previous case are all ≥ 1 . Thus we either obtain N independent functions in $\leq N$ stages, or we arrive at the present situation in $\leq N$ stages.] We now have $N' = N_1 + \dots + N_\mu < N$ independent functions. If α exists, then it must satisfy the N' equations

$$(*) \quad \begin{cases} C_{jk}^i = \bar{C}_{jk}^i \circ \alpha \\ \vdots \\ C_{jk,l_1, \dots, l_\mu}^i = \bar{C}_{jk,l_1, \dots, l_\mu}^i \circ \alpha. \end{cases}$$

In the same way that we obtained equations (4), we can use equations (*) to deduce N' linear equations for $\omega^1 - \bar{\omega}^1, \dots, \omega^N - \bar{\omega}^N$. Moreover, the rank of the matrix for these equations is N' , so we can solve for $N - N'$ of the unknowns in terms of the other N' . Without loss of generality, we can assume that these equations can be solved for the last $N - N'$ of the $\omega^i - \bar{\omega}^i$ in terms of the first N' of the $\omega^i - \bar{\omega}^i$. Then clearly the diffeomorphism α has the desired properties if it satisfies (*) as well as

$$(**) \quad \omega^i = \alpha^* \bar{\omega}^i, \quad i = 1, \dots, N'.$$

We claim that there are always such diffeomorphisms α , in fact, an $(N - N')$ -parameter family of them. To prove this, we look for the graph of α , as a subset

of $M \times \bar{M}$. Let $\pi: M \times \bar{M} \rightarrow M$ and $\bar{\pi}: M \times \bar{M} \rightarrow \bar{M}$ be the projections. Since the C_{jk}^i, \dots and \bar{C}_{jk}^i, \dots in (*) are independent functions, the set

$$\mathcal{M} = \{x \in M \times \bar{M} : C_{ij} \circ \pi(x) = \bar{C}_{jk}^i \circ \bar{\pi}(x), \dots\}$$

is a submanifold, of dimension $2N - N'$. We will denote the restrictions of $\pi^*\omega^i$ and $\bar{\pi}^*\bar{\omega}^i$ to \mathcal{M} simply by $\pi^*\omega^i$ and $\bar{\pi}^*\bar{\omega}^i$. Consider the ideal \mathcal{I} of forms on \mathcal{M} generated by the 1-forms

$$\pi^*\omega^i - \bar{\pi}^*\bar{\omega}^i, \quad i = 1, \dots, N'.$$

We have

$$\begin{aligned} d(\pi^*\omega^i - \bar{\pi}^*\bar{\omega}^i) &= \sum (C_{jk}^i \circ \pi) \pi^*\omega^j \wedge \pi^*\omega^k - \sum (\bar{C}_{jk}^i \circ \bar{\pi}) \bar{\pi}^*\bar{\omega}^j \wedge \bar{\pi}^*\bar{\omega}^k \\ &= \sum (C_{jk}^i \circ \pi) [\pi^*\omega^j \wedge \pi^*\omega^k - \bar{\pi}^*\bar{\omega}^j \wedge \bar{\pi}^*\bar{\omega}^k] \quad (\text{on } \mathcal{M}) \\ &= \sum (C_{jk}^i \circ \pi) [(\pi^*\omega^j \wedge \bar{\pi}^*\bar{\omega}^i) \wedge \pi^*\omega^k \\ &\quad - \bar{\pi}^*\bar{\omega}^i \wedge (\pi^*\omega^j - \bar{\pi}^*\bar{\omega}^j)], \end{aligned}$$

which is in \mathcal{I} . Thus there is a submanifold \mathcal{M}' of \mathcal{M} on which the forms $\pi^*\omega^i - \bar{\pi}^*\bar{\omega}^i$ all vanish. This submanifold \mathcal{M}' has dimension $(2N - N') - N' = 2(N - N')$. There is an $(N - N')$ -parameter family of N -dimensional submanifolds \mathcal{M}'' of \mathcal{M}' which project one-one onto M . Each of these is the graph of an appropriate α .

Finally, let us return to the case of two Riemannian manifolds M^n and \bar{M}^n . We immediately pass to the principal bundles $O(TM)$ and $O(T\bar{M})$ of orthonormal frames. On these bundles we have forms $\theta = (\theta^i)$, $\omega = (\omega_j^i)$ and $\bar{\theta} = (\bar{\theta}^i)$, $\bar{\omega} = (\bar{\omega}_j^i)$. Recall that for $u = (u_1, \dots, u_n) \in O(TM)$ and a tangent vector $Y \in O(TM)_u$, we have

$$\pi_* Y_u = \sum_{i=1}^n \theta^i(Y_u) \cdot u_i,$$

where $\pi: O(TM) \rightarrow M$ is the projection map. In particular, $\theta(Y_u) = 0$ if and only if $\pi_* Y_u = 0$. Now any isometry $\alpha: M \rightarrow \bar{M}$ gives rise to a diffeomorphism $\tilde{\alpha}: O(TM) \rightarrow O(T\bar{M})$, and $\tilde{\alpha}^*\bar{\theta} = \theta$, $\tilde{\alpha}^*\bar{\omega} = \omega$. Conversely, suppose we have a diffeomorphism $\beta: O(TM) \rightarrow O(T\bar{M})$ with $\beta^*\bar{\theta} = \theta$. If c is any curve in the fibre of $O(TM)$ at p , then for all t we have $\pi_* c'(t) = 0$, and thus

$$\begin{aligned} 0 &= \theta(c'(t)) = \beta^*\bar{\theta}(c'(t)) \\ &= \bar{\theta}(\beta_* c'(t)) \\ \implies 0 &= \bar{\pi}_* \beta_* c'(t) = (\bar{\pi} \circ \beta \circ c)'(t). \end{aligned}$$

Thus $\bar{\pi} \circ \beta \circ c$ is constant. This shows that β takes fibres to fibres, so there is a map $\alpha: M \rightarrow \bar{M}$ with $\bar{\pi} \circ \beta = \alpha \circ \pi$. Moreover, if $u = (u_1, \dots, u_n) \in O(TM)$ and $Y_u \in O(TM)_u$ satisfies $\pi_* Y_u = u_j$, then $\theta^i(Y_u) = \delta_j^i$, so

$$\begin{aligned} \delta_j^i &= \beta^* \bar{\theta}^i(Y_u) \\ &= \bar{\theta}^i(\beta_* Y_u) \\ &= i^{\text{th}} \text{ component of } \bar{\pi}_* \beta_* Y_u \text{ with respect to } \beta(u) \\ &= \text{''} \quad \text{''} \quad \text{''} \alpha_* \pi_* Y_u \quad \text{''} \quad \text{''} \quad \text{''} \beta(u) \\ &= \text{''} \quad \text{''} \quad \text{''} \alpha_*(u_j) \quad \text{''} \quad \text{''} \quad \text{''} \beta(u). \end{aligned}$$

Thus β must be

$$\beta(u) = (\alpha_* u_1, \dots, \alpha_* u_n),$$

i.e., $\beta = \tilde{\alpha}$. In particular, α is an isometry. Thus we see that the existence of an isometry $\alpha: M \rightarrow \bar{M}$ is equivalent to the existence of a diffeomorphism $\beta: O(TM) \rightarrow O(T\bar{M})$ such that $\beta^* \bar{\theta}^i = \theta^i$. Hence it is also equivalent to the existence of a diffeomorphism $\beta: O(TM) \rightarrow O(T\bar{M})$ such that $\beta^* \bar{\theta}^i = \theta^i$ and $\beta^* \bar{\omega}_j^i = \omega_j^i$. We have just seen how to decide whether such a β exists, since the θ^i and ω_j^i are everywhere linearly independent and span the 1-forms, and similarly for the $\bar{\theta}^i$ and $\bar{\omega}_j^i$. The first step is to compute the $d\theta^i$ and $d\omega_j^i$ in terms of the θ^i and ω_j^i . We already have the structural equations (pg. II.329),

$$(1) \quad d\theta^i = - \sum_j \omega_j^i \wedge \theta^j$$

$$\begin{aligned} (2) \quad d\omega_j^i &= - \sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i \\ &= - \sum_k \omega_k^i \wedge \omega_j^k - \sum_{k < l} A_{ijkl} \theta^k \wedge \theta^l, \quad \text{say.} \end{aligned}$$

These functions A_{ijkl} are the first set which we have to examine. Now if $s = (X_1, \dots, X_n)$ is an orthonormal moving frame, then its dual forms and connection forms are $\theta^i = s^* \bar{\theta}^i$ and $\omega_j^i = s^* \bar{\omega}_j^i$. So s^* of equation (2) gives

$$d\omega_j^i = - \sum_k \omega_k^i \wedge \omega_j^k - \sum_{k < l} (A_{ijkl} \circ s) \theta^k \wedge \theta^l.$$

On the other hand, we have

$$\begin{aligned} d\omega_j^i &= - \sum_k \omega_k^i \wedge \omega_j^k + \Omega_j^i \\ &= - \sum_k \omega_k^i \wedge \omega_j^k + \sum_{k < l} \langle R(X_k, X_l) X_j, X_i \rangle \theta^k \wedge \theta^l. \end{aligned}$$

Thus we see that

$$A_{ijkl}(u) = -\langle R(u_k, u_l)u_j, u_i \rangle = \mathcal{R}(u_i, u_j, u_k, u_l).$$

The next set of functions which we need to look at are those appearing in the expansion

$$dA_{ijkl} = \sum_{\mu, \nu} (\quad) \omega_\nu^\mu + \sum_{\mu} A_{ijkl, \mu} \theta^\mu.$$

Taking s^* of this equation, and evaluating at a tangent vector X of M , we get

$$\begin{aligned} X(\mathcal{R}(X_i, X_j, X_k, X_l)) &= X(A_{ijkl} \circ s) = d(A_{ijkl} \circ s)(X) = s^*(dA_{ijkl})(X) \\ &= \sum_{\mu, \nu} [(\quad) \circ s] \omega_\nu^\mu(X) + \sum_{\mu} [A_{ijkl, \mu} \circ s] \theta^\mu(X). \end{aligned}$$

Since

$$\begin{aligned} X(\mathcal{R}(X_i, X_j, X_k, X_l)) &= (\nabla \mathcal{R})(X_i, X_j, X_k, X_l, X) \\ &\quad + \mathcal{R}(\nabla_X X_i, \dots) + \mathcal{R}(X_i, \nabla_X X_j, \dots) + \dots, \end{aligned}$$

we see that we must have

$$dA_{ijkl} = \sum_{\mu} A_{\mu jkl} \omega_i^\mu + \dots + \sum_{\mu} A_{ijk\mu} \omega_l^\mu + \sum_{\mu} A_{ijkl, \mu} \theta^\mu,$$

where

$$A_{ijkl, \mu}(u) = (\nabla \mathcal{R})(u_i, u_j, u_k, u_l, u_\mu).$$

Similarly, we get

$$dA_{ijkl, \mu} = \sum_{\nu} A_{\nu jkl, \mu} \omega_i^\nu + \dots + \sum_{\nu} A_{ijk\nu, \mu} \omega_l^\nu + \sum_{\nu} A_{ijkl, \nu} \omega_\mu^\nu + \sum_{\nu} A_{ijkl, \mu\nu} \theta^\nu,$$

where

$$A_{ijkl, \mu\nu}(u) = (\nabla \nabla \mathcal{R})(u_i, u_j, u_k, u_l, u_\mu, u_\nu),$$

and so on. So we see that after computing a finite number of the functions $\mathcal{R}, \nabla \mathcal{R}, \nabla \nabla \mathcal{R}, \dots$ we can finally decide if the desired isometry α exists (provided we can keep track of what in the world we are doing, which doesn't seem very likely).

ADDENDUM 4

BETTER IMBEDDING INVARIANTS

There is a theory, due to Burstin, Mayer, and Allendoerfer, which shows that certain tensors are a complete set of invariants for submanifolds $M^n \subset N^m$ of a manifold N of constant curvature. (As in the theory of curves, we have to impose certain conditions on M , which say, roughly speaking, that at each point M bends in the same number of directions.) One almost never sees any applications of this theory nowadays, but perhaps that is partly because the classical expositions make it so inaccessible. In our presentation, we will first consider a special case of the general problem, so as not to be overwhelmed with details at the beginning.

For a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, the “Fundamental Lemma of Riemannian geometry” tells us that there is a unique connection ∇ on TM which is compatible with the metric and also symmetric. The following Lemma gives, under certain conditions, an analogous characterization of the normal connection D on the normal bundle $\text{Nor } M$ of M in N .

65. LEMMA. Let $M \subset N$ have normal bundle $\text{Nor } M$ and second fundamental form s . Suppose that $s: M_p \times M_p \rightarrow M_p^\perp$ is onto for all p . Then the normal connection D in $\text{Nor } M$ is the unique connection δ such that

(1) δ is compatible with the metric in $\text{Nor } M$:

$$X(\langle \xi, \eta \rangle) = \langle \delta_X \xi, \eta \rangle + \langle \xi, \delta_X \eta \rangle \quad \text{for sections } \xi, \eta \text{ of } \text{Nor } M,$$

(2) δ satisfies the Codazzi-Mainardi equations:

$$\begin{aligned} \perp(R'(X, Y)Z) &= [\delta_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z)] \\ &\quad - [\delta_Y(s(X, Z)) - s(\nabla_Y X, Z) - s(X, \nabla_Y Z)]. \end{aligned}$$

PROOF. Consider the expression

$$\begin{aligned} &\langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle - \langle \delta_{Y_1} s(X, Y_2), s(Z_1, Z_2) \rangle \\ &\quad - \langle \delta_{Y_1} s(Z_1, Z_2), s(X, Y_2) \rangle + \langle \delta_{Z_1} s(Y_1, Z_2), s(X, Y_2) \rangle \\ &\quad + \langle \delta_{Z_1} s(X, Y_2), s(Y_1, Z_2) \rangle - \langle \delta_X s(Z_1, Y_2), s(Y_1, Z_2) \rangle. \end{aligned}$$

Condition (2) shows that each row can be expressed in terms of the vector fields X, Y_i, Z_i . But condition (1) shows that the sum of the two terms involving δ_{Y_1} or δ_{Z_1} can also be expressed in this way. Thus we can write

$$(3) \quad \langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle - \langle \delta_X s(Z_1, Y_2), s(Y_1, Z_2) \rangle = \dots,$$

where ... can be expressed in terms of the vector fields. So we have as well

$$(3') \quad \langle \delta_X s(Y_2, Z_1), s(Z_2, Y_1) \rangle - \langle \delta_X s(Z_2, Z_1), s(Y_2, Y_1) \rangle = \dots$$

Adding (3) and (3'), we obtain

$$(4) \quad \langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle - \langle \delta_X s(Z_1, Z_2), s(Y_1, Y_2) \rangle = \dots$$

But by (1) we also have

$$(5) \quad \langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle + \langle \delta_X s(Z_1, Z_2), s(Y_1, Y_2) \rangle = \dots$$

So by adding (4) and (5) we obtain

$$\langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle = \dots$$

Since $s: M_p \times M_p \rightarrow M_p^\perp$ is onto, this shows that $\delta_X s(Y_1, Y_2)$ is uniquely determined by X_p, Y_1, Y_2 . Since every section of $\text{Nor } M$ is a linear combination, over the C^∞ functions, of sections of the form $s(Y_1, Y_2)$, this shows that δ is uniquely determined. ♦

Remark: Naturally we are mainly interested in the case where N has constant curvature, in which case the left side of (2) is 0. Now given any bundle $\varpi: E \rightarrow M$ with a metric $\langle \cdot, \cdot \rangle$, and a symmetric section s of $\text{Hom}(TM \times TM, E)$, we can consider the “Codazzi-Mainardi equations”

$$\begin{aligned} [\delta_X(s(Y, Z)) - s(\nabla_X Y, Z) - s(Y, \nabla_X Z)] = \\ [\delta_Y(s(X, Z)) - s(\nabla_Y X, Z) - s(X, \nabla_Y Z)]. \end{aligned}$$

The proof of Lemma 65 shows that if $s: M_p \times M_p \rightarrow \varpi^{-1}(p)$ is always onto, then there is at most one δ compatible with $\langle \cdot, \cdot \rangle$ which satisfies this equation. However, there may not be any such δ (unless s is always one-one).

Now for a submanifold $M \subset N$ we will denote the induced metric $\langle \cdot, \cdot \rangle$ on M by \mathcal{F}_0 , and define a tensor \mathcal{F}_1 by

$$\mathcal{F}_1(X_1, X_2, Y_1, Y_2) = \langle s(X_1, X_2), s(Y_1, Y_2) \rangle.$$

66. PROPOSITION. Let $M, \bar{M} \subset N$ be connected submanifolds of a complete simply-connected manifold N of constant curvature. Suppose that the second fundamental forms $s: M_p \times M_p \rightarrow M_p^\perp$ and $\bar{s}: \bar{M}_q \times \bar{M}_q \rightarrow \bar{M}_q^\perp$ are onto at all points. Let $\phi: M \rightarrow \bar{M}$ be a diffeomorphism such that

$$\phi^* \bar{\mathcal{F}}_0 = \mathcal{F}_0 \quad \text{and} \quad \phi^* \bar{\mathcal{F}}_1 = \mathcal{F}_1.$$

Then ϕ is the restriction of an isometry of N .

PROOF. First of all, since $\phi^* \bar{\mathcal{F}}_0 = \mathcal{F}_0$, we have

$$(1) \quad \phi_*(\nabla_X Y) = \bar{\nabla}_{\phi_* X} \phi_* Y.$$

Now let $\{X_\alpha\}$ be any set of vectors in M_p which span M_p , and let $\bar{X}_\alpha = \phi_*(X_\alpha) \in \bar{M}_{f(p)}$. Consider the vectors $s(X_\alpha, X_\beta) \in M_p^\perp$, and the corresponding vectors $\bar{s}(\bar{X}_\alpha, \bar{X}_\beta) \in \bar{M}_{f(p)}^\perp$. By hypothesis, we have

$$\langle s(X_\alpha, X_\beta), s(X_\gamma, X_\delta) \rangle = \langle \bar{s}(\bar{X}_\alpha, \bar{X}_\beta), \bar{s}(\bar{X}_\gamma, \bar{X}_\delta) \rangle.$$

Since the second fundamental forms are onto M_p^\perp and $\bar{M}_{f(p)}^\perp$, this implies (Problem 25) that there is a unique inner product preserving isomorphism $M_p^\perp \rightarrow \bar{M}_{f(p)}^\perp$ which takes $s(X_\alpha, X_\beta)$ to $\bar{s}(\bar{X}_\alpha, \bar{X}_\beta)$. This isomorphism cannot depend on the $\{X_\alpha\}$, for if we also have spanning vectors $\{Y_\alpha\}$, we can consider the set $\{X_\alpha\} \cup \{Y_\alpha\}$.

By applying this construction for all $p \in M$, we obtain a bundle isomorphism $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ covering ϕ such that $\tilde{\phi}$ preserves inner products and second fundamental forms:

$$(2) \quad \langle \tilde{\phi}(\xi), \tilde{\phi}(\eta) \rangle = \langle \xi, \eta \rangle \quad \text{for sections } \xi, \eta \text{ of } \text{Nor } M$$

$$(3) \quad \tilde{\phi}(s(X, Y)) = \bar{s}(\phi_* X, \phi_* Y).$$

We claim that $\tilde{\phi}$ also preserves the normal connections:

$$(4) \quad \tilde{\phi}(D_X \xi) = \bar{D}_{\phi_* X}(\tilde{\phi}(\xi)).$$

To prove this we note that since every section of $\text{Nor } \bar{M}$ is uniquely of the form $\tilde{\phi}(\xi)$, and every tangent vector of \bar{M} is uniquely of the form $\phi_* X$, we can define a connection δ on $\text{Nor } \bar{M}$ with

$$\delta_{\phi_* X}(\tilde{\phi}(\xi)) = \tilde{\phi}(D_X \xi).$$

Now the connection D is compatible with the metric and satisfies the Codazzi-Mainardi equations; applying the equations (1)–(3), we find that δ is compatible with the metric and satisfies the Codazzi-Mainardi equations (for \bar{M}). Hence $\delta = \bar{D}$, by Lemma 65. This proves (4).

The desired result now follows from Theorem 20. ♦

When we have a manifold $M \subset N$ whose second fundamental form does not fill up the normal bundle, we will have to differentiate more times, precisely as in the case of curves. Notice that the subspace of N_p spanned by M_p and

$s_p(M_p \times M_p) \subset M_p^\perp$ can also be described as the space spanned by all X_p and $\nabla'_{X_p} Y$ for vector fields X, Y on M . But we can also consider $\nabla'_{X_p}(\nabla'_Y Z)$, etc., and thereby obtain more vectors in N_p . To simplify the notation, we will write

$$\begin{aligned}\nabla'(X, Y) &= \nabla'_X Y \\ \nabla'(X, Y, Z) &= \nabla'_X(\nabla'_Y Z), \quad \text{etc.}\end{aligned}$$

We define the k^{th} **osculating space** $\text{Osc}^k M_p \subset N_p$ of M at p to be the subspace of N_p which is spanned by all

$$X_1(p), \nabla'(X_1, X_2)(p), \dots, \nabla'(X_1, \dots, X_k)(p),$$

for vector fields X_1, \dots, X_k on M . Thus the 1st osculating space $\text{Osc}^1 M_p$ is just M_p . It will also be convenient to define $\text{Osc}^0 M_p$ to be the $\{0\}$ subspace of M_p .

A submanifold $M \subset N$ will be called **nicely curved** if the dimension of each osculating space $\text{Osc}^k M_p$ is the same for all $p \in M$. (A curve c in N is nicely curved if and only if it has the property that if some curvature function κ_k is non-zero at one point, then κ_k is non-zero everywhere.) Henceforth we will consider only nicely curved submanifolds $M \subset N$. It is easy to see that for each k we then have a vector bundle $\text{Osc}^k M$ over M , whose fibre over p is $\text{Osc}^k M_p$. If ξ is a smooth section of $\text{Osc}^k M$, then ξ is locally a sum of terms $f \cdot \nabla'(X_1, \dots, X_r)$ for smooth f and X_i , and $r \leq k$. Since

$$\nabla'_X(f \nabla'(X_1, \dots, X_r)) = X(f) \cdot \nabla'(X_1, \dots, X_r) + f \cdot \nabla'(X, X_1, \dots, X_r),$$

we see that

$$(*) \quad \xi \text{ a section of } \text{Osc}^k M \implies \nabla'_{X_p} \xi \in \text{Osc}^{k+1} M_p.$$

It is easy to see that if $\text{Osc}^k M = \text{Osc}^{k+1} M$, then also $\text{Osc}^{k+1} M = \text{Osc}^{k+2} M = \dots$. So there is some $\ell \geq 1$ with

$$\text{Osc}^0 M \subsetneq \text{Osc}^1 M \subsetneq \dots \subsetneq \text{Osc}^\ell M = \text{Osc}^{\ell+1} M = \dots$$

The letter ℓ will always have this significance. Notice that $\text{Osc}^\ell M_p$ need not be all of N_p ; the dimension of $\text{Osc}^\ell M_p$ (for any $p \in M$) will be called the **formal imbedding number** $\#(M)$ of M .

67. PROPOSITION. If $M \subset N$ is nicely curved, then the distribution $p \mapsto \text{Osc}^\ell M_p$ on M is parallel along every curve in M (as defined on page 28).

Consequently, if N is a manifold of constant curvature, and M is connected, then M is contained in some $\#(M)$ -dimensional totally geodesic submanifold of N (but not in any lower dimensional totally geodesic submanifold).

PROOF. To prove the first part, it obviously suffices to work locally. In a neighborhood U of any point $p \in M$ we can choose smooth linearly independent sections $\xi_1, \dots, \xi_{\#(M)}$ of $\text{Osc}^\ell M$. For a curve c in U , let V_μ be the vector field along c given by $V_\mu(t) = \xi_\mu(c(t))$. Then $(*)$ says that

$$\frac{D'V_\mu}{dt} \in \text{Osc}^{l+1} M_{c(t)} = \text{Osc}^l M_{c(t)};$$

thus there are smooth functions $f_{\mu\nu}$ such that

$$\frac{D'V_\mu}{dt} = \sum_\nu f_{\mu\nu} V_\nu.$$

Now let W be any vector field along c with $D'W/dt = 0$. Then

$$\begin{aligned} \frac{d}{dt} \langle W, V_\mu \rangle &= \left\langle \frac{D'W}{dt}, V_\mu \right\rangle + \left\langle W, \frac{D'V_\mu}{dt} \right\rangle \\ &= \langle 0, V_\mu \rangle + \left\langle W, \sum_\nu f_{\mu\nu} V_\nu \right\rangle \\ &= \sum_\nu f_{\mu\nu} \langle W, V_\nu \rangle. \end{aligned}$$

This is a system of differential equations for the functions $\langle W, V_\mu \rangle$. One solution is $\langle W, V_\mu \rangle = 0$ for all μ . So by uniqueness of solutions we see that if $W(0)$ is perpendicular to $\text{Osc}^\ell M_{c(0)}$, then $W(t)$ is perpendicular to $\text{Osc}^\ell M_{c(t)}$ for all t . This proves that $\text{Osc}^\ell M$ is parallel along c .

The second part follows from Corollary 11. ♦

We now define the k^{th} **normal space** $\text{Nor}^k M_p$ of M at p to be the orthogonal complement of $\text{Osc}^k M_p$ in $\text{Osc}^{k+1} M_p$. Thus we have

$$\text{Osc}^{k+1} M_p = \text{Osc}^k M_p \oplus \text{Nor}^k M_p.$$

Notice that $\text{Nor}^0 M_p$ is just M_p , while $\text{Nor}^k M_p$ has dimension 0 for $k \geq \ell - 1$. It will also be convenient to let $\text{Nor}^{-1} M_p$ be the $\{0\}$ subspace of M_p . Each $\text{Nor}^k M_p \subset N_p$ has an orthogonal complement in N_p , and thus we have two projections

$$\mathbb{T}^k : N_p \rightarrow \text{Nor}^k M_p$$

$$\mathbb{L}^k : N_p \rightarrow \text{orthogonal complement of } \text{Nor}^k M_p \text{ in } N_p.$$

Notice that $\mathsf{T}^0 = \mathsf{T}: N_p \rightarrow M_p$ and $\mathsf{L}^0 = \mathsf{L}: N_p \rightarrow M_p^\perp$ (but note that T^k goes into a subspace of M_p^\perp for $k > 0$). We shall actually use only the projections T^k .

For nicely curved $M \subset N$ we clearly have, for each k , a vector bundle $\text{Nor}^k M$ whose fibre at p is $\text{Nor}^k M_p$. The bundle $\text{Nor}^0 M$ is just the tangent bundle TM , while the bundles $\text{Nor}^k M$ for $k > 0$ are all subbundles of the normal bundle $\text{Nor} M$. There are natural Riemannian metrics $\langle \cdot, \cdot \rangle$ on all bundles $\text{Nor}^k M$, since they are all subbundles of $(TN)|_M$.

The 1st normal space $\text{Nor}^1 M_p$ is the subspace of N_p spanned by all $s(X_p, Y_p)$ for $X_p, Y_p \in M_p$. In general, given vector fields X_1, \dots, X_{k+1} on M , consider

$$\mathsf{T}^k(\nabla'(X_1, \dots, X_{k+1})).$$

It is easily checked that this expression is linear in each X_i over the C^∞ functions (compare pg. III.4). So its value at p depends only on the values of the X_i at p and we can define

$$s^k(X_{1p}, \dots, X_{k+1p}) = \mathsf{T}^k(\nabla'(X_1, \dots, X_{k+1})(p)) \in \text{Nor}^k M_p$$

for any vector fields X_i extending X_{ip} . Clearly $\text{Nor}^k M_p$ is spanned by the image of s^k . It seems reasonable to let s^0 denote the identity map of M_p into $\text{Nor}^0 M_p = M_p$.

68. LEMMA. If $M \subset N$, where N has constant curvature, then s^k is symmetric.

PROOF. First we have

$$\begin{aligned} \nabla'(X_1, \dots, X_k, X_{k+1})(p) - \nabla'(X_1, \dots, X_{k+1}, X_k)(p) \\ = \nabla'(X_1, \dots, X_{k-1}, [X_k, X_{k+1}])(p) \in \text{Osc}^k M_p, \end{aligned}$$

so T^k of the left side is 0, which proves that s^k is symmetric in X_k and X_{k+1} . We also have, for example,

$$\begin{aligned} \nabla'(X_1, \dots, X_{k-1}, X_k, X_{k+1})(p) - \nabla'(X_1, \dots, X_k, X_{k-1}, X_{k+1})(p) \\ = \nabla'(X_1, \dots, X_{k-2}, \nabla'_{X_{k-1}}(\nabla'_{X_k} X_{k+1}) - \nabla'_{X_k}(\nabla'_{X_{k-1}} X_k))(p) \\ = \nabla'(X_1, \dots, X_{k-2}, \nabla'_{[X_{k-1}, X_k]} X_{k+1} + R'(X_{k-1}, X_k) X_{k+1})(p). \end{aligned}$$

Since $R'(X_{k-1}, X_k) X_{k+1}$ is tangent to M , this is in $\text{Osc}^{k-1} M_p$, so again T^k of the left side is 0. Similarly, s^k is symmetric under interchange of any two adjacent arguments. ♦

For vector fields X_1, \dots, X_{k+1} on a nicely curved submanifold $M \subset N$ we can write

$$s^k(X_1, \dots, X_{k+1}) = \nabla'(X_1, \dots, X_{k+1}) + \xi, \quad \xi \text{ a section of } \text{Osc}^k M.$$

Then

$$\mathbb{T}^{k+1} \nabla'_{X_p} s^k(X_1, \dots, X_{k+1}) = \mathbb{T}^{k+1} \nabla'_{X_p} \nabla'(X_1, \dots, X_{k+1}),$$

since $\nabla'_{X_p} \xi \in \text{Osc}^{k+1} M_p$ by (*), on page 166. Thus we have

$$(**) \quad \mathbb{T}^{k+1} \nabla'_{X_p} s^k(X_1, \dots, X_{k+1}) = s^{k+1}(X_p, X_1(p), \dots, X_{k+1}(p)).$$

Now suppose we have vector fields $\{Y_\alpha\}$ which span the tangent space of M in a neighborhood of p . Every element of $\text{Nor}^1 M_p$, for example, can be written as

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha(p), Y_\beta(p)).$$

This expression is usually not unique (even if the $Y_\alpha(p)$ are linearly independent). But suppose that we have constants $c_{\alpha\beta}$ with

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha(p), Y_\beta(p)) = 0.$$

Let \mathcal{J} be a collection of pairs (α, β) such that

$$\{s(X_\alpha(p), X_\beta(p)) : (\alpha, \beta) \in \mathcal{J}\}$$

is a basis of $\text{Nor}^1 M_p$. Since M is nicely curved, it follows that $\{s(X_\alpha(q), X_\beta(q)) : (\alpha, \beta) \in \mathcal{J}\}$ is a basis of $\text{Nor}^1 M_q$ for all points q in a neighborhood of p . Now consider the section

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha, Y_\beta)$$

of $\text{Nor}^1 M$, where the $c_{\alpha\beta}$ denote constant functions. In a neighborhood of p we can write

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha, Y_\beta) = \sum_{(\alpha, \beta) \in \mathcal{J}} f_{\alpha\beta} \cdot s(Y_\alpha, Y_\beta)$$

for *unique* smooth functions $f_{\alpha\beta}$. Clearly $f_{\alpha\beta}(p) = 0$. Applying ∇'_{X_p} to the above equation we thus obtain

$$\begin{aligned} \sum c_{\alpha\beta} \cdot \nabla'_{X_p} s(Y_\alpha, Y_\beta) &= \sum_{(\alpha, \beta) \in \mathcal{J}} X_p(f_{\alpha\beta}) \cdot s(Y_\alpha(p), Y_\beta(p)) + 0 \\ &\in \text{Nor}^1 M_p. \end{aligned}$$

Consequently,

$$0 = \sum c_{\alpha\beta} \cdot \mathbb{T}^2 \nabla'_{X_p} s(Y_\alpha, Y_\beta) = \sum c_{\alpha\beta} \cdot s(X_p, Y_\alpha(p), Y_\beta(p)) \quad \text{by (**).}$$

Thus we see that

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha(p), Y_\beta(p)) = 0 \implies \sum c_{\alpha\beta} \cdot s(X_p, Y_\alpha(p), Y_\beta(p)) = 0.$$

It follows that there is a well-defined map from $\text{Nor}^1 M_p$ to $\text{Nor}^2 M_p$ under which

$$\sum c_{\alpha\beta} \cdot s(Y_\alpha(p), Y_\beta(p)) \mapsto \sum c_{\alpha\beta} \cdot s(X_p, Y_\alpha(p), Y_\beta(p)).$$

This map doesn't depend on the choice of $\{Y_\alpha\}$, for if we also have spanning vector fields $\{Z_\alpha\}$, we can apply the above argument to the collection $\{Y_\alpha\} \cup \{Z_\alpha\}$. The argument clearly works for all k , so we see that there is a well-defined bilinear map

$$\mathbf{s}^k : M_p \times \text{Nor}^k M_p \rightarrow \text{Nor}^{k+1} M_p$$

such that

$$\mathbf{s}^k(X_p, \mathbf{s}^k(X_{1p}, \dots, X_{k+1p})) = \mathbf{s}^{k+1}(X_p, X_{1p}, \dots, X_{k+1p}).$$

Now suppose we have any section ξ of $\text{Nor}^k M$. Locally ξ can be written as a sum of terms $f \cdot \mathbf{s}^k(X_1, \dots, X_{k+1})$. Since

$$\begin{aligned} \nabla'_{X_p}(f \cdot \mathbf{s}^k(X_1, \dots, X_{k+1})) &= X_p(f) \cdot \mathbf{s}^k(X_{1p}, \dots, X_{k+1p}) \\ &\quad + f(p) \cdot \nabla'_{X_p} \mathbf{s}^k(X_1, \dots, X_{k+1}), \end{aligned}$$

we see that

$$\mathbb{T}^{k+1} \nabla'_{X_p}(f \cdot \mathbf{s}^k(X_1, \dots, X_{k+1})) = f(p) \cdot \mathbb{T}^{k+1} \nabla'_{X_p} \mathbf{s}^k(X_1, \dots, X_{k+1}).$$

Then (**) shows that

$$(***) \quad \mathbb{T}^{k+1} \nabla'_{X_p} \xi = \mathbf{s}^k(X_p, \xi_p), \quad \xi \text{ a section of } \text{Nor}^k M.$$

Now we will consider all the other components of $\nabla'_{X_p} \xi$ for ξ a section of $\text{Nor}^k M$. First we note

69. LEMMA. If ξ is a section of $\text{Nor}^k M$ and $X_p \in M_p$, then

$$\nabla'_{X_p} \xi \in \text{Nor}^{k-1} M_p \oplus \text{Nor}^k M_p \oplus \text{Nor}^{k+1} M_p.$$

PROOF. Since ξ is a section of $\text{Osc}^{k+1} M$, we have $\nabla'_{X_p} \xi \in \text{Osc}^{k+2} M_p$. Now if η is any section of $\text{Osc}^j M$ for $j < k$, then $\langle \xi, \eta \rangle = 0$, so

$$0 = X_p(\langle \xi, \eta \rangle) = \langle \nabla'_{X_p} \xi, \eta(p) \rangle + \langle \xi(p), \nabla'_{X_p} \eta \rangle.$$

If we also have $j < k - 1$, then $\nabla'_{X_p} \eta \in \text{Osc}^k M_p$, so $\langle \xi(p), \nabla'_{X_p} \eta \rangle = 0$, and hence $\langle \nabla'_{X_p} \xi, \eta(p) \rangle = 0$. ♦

Thus we see that for a section ξ of $\text{Nor}^k M$ we can write

$$\nabla'_{X_p} \xi = \mathbb{T}^{k-1}(\nabla'_{X_p} \xi) + \mathbb{T}^k(\nabla'_{X_p} \xi) + \mathbb{T}^{k+1}(\nabla'_{X_p} \xi).$$

The third term of this decomposition is already given by (**). For the first term we have a result which generalizes Proposition 12.

70. PROPOSITION. If ξ is a section of $\text{Nor}^k M$ and $X_p \in M_p$, then the vector $\mathbb{T}^{k-1}(\nabla'_{X_p} \xi) \in \text{Nor}^{k-1} M_p$ satisfies

$$\begin{aligned} \langle \mathbb{T}^{k-1}(\nabla'_{X_p} \xi), \eta_p \rangle &= \langle \nabla'_{X_p} \xi, \eta_p \rangle = -\langle \xi(p), \mathbf{s}^{k-1}(X_p, \eta_p) \rangle \\ &\text{for all } \eta_p \in \text{Nor}^{k-1} M_p. \end{aligned}$$

Consequently, $\mathbb{T}^{k-1}(\nabla'_{X_p} \xi)$ depends only on X_p and ξ_p .

PROOF. If η is a section of $\text{Nor}^{k-1} M$ extending η_p , then $\langle \xi, \eta \rangle = 0$, so

$$\begin{aligned} 0 &= X_p(\langle \xi, \eta \rangle) = \langle \nabla'_{X_p} \xi, \eta_p \rangle + \langle \xi(p), \nabla'_{X_p} \eta \rangle \\ &= \langle \nabla'_{X_p} \xi, \eta_p \rangle + \langle \xi(p), \mathbb{T}^k \nabla'_{X_p} \eta \rangle, \end{aligned}$$

since $\xi(p) \in \text{Nor}^k M_p$, by assumption. Now apply (**). ♦

For any vector $\xi_p \in \text{Nor}^k M_p$ we can now let

$$A_{\xi_p}^k(X_p) = -\mathbb{T}^{k-1}(\nabla'_{X_p} \xi),$$

for any section ξ of $\text{Nor}^k M$ extending ξ_p , so that we have a map

$$A_{\xi_p}^k : M_p \rightarrow \text{Nor}^{k-1} M_p$$

satisfying

$$\langle A_{\xi_p}^k(X_p), \eta_p \rangle = \langle \xi_p, s^{k-1}(X_p, \eta_p) \rangle.$$

(Note that for $k = 0$ we are dealing with the 0 map.) For convenience, we will sometimes write

$$A^k(\xi_p; X_p) \quad \text{for} \quad A_{\xi_p}^k(X_p).$$

Finally, for the expression $\mathsf{T}^k(\nabla'_{X_p}\xi)$ we introduce a new symbol,

$$D_{X_p}^k \xi = \mathsf{T}^k(\nabla'_{X_p}\xi), \quad \xi \text{ a section of } \text{Nor}^k M.$$

It is easy to check that D^k is a connection on $\text{Nor}^k M$ which is compatible with the metric $\langle \cdot, \cdot \rangle$ on $\text{Nor}^k M$. We can now write our decomposition of $\nabla'_{X_p}\xi$ as

The Frenet Equations:

$$\begin{aligned} \nabla'_{X_p}\xi &= -A_{\xi_p}^k(X_p) + D_{X_p}^k \xi + s^k(X_p, \xi_p), \\ \text{for } \xi &\text{ a section of } \text{Nor}^k M \text{ and } X_p \in M_p. \end{aligned}$$

The terms $A_{\xi_p}^k(X_p)$ and $s^k(X_p, \xi_p)$ are completely determined by the maps s^{k-1} , s^k , and s^{k+1} . These Frenet equations essentially contain the Frenet equations for a curve when M is 1-dimensional; in general, they contain the Gauss equations (for $k = 0$) and (part of) the Weingarten equations for $k = 1$.

71. THEOREM. Let $M^n, \bar{M}^n \subset N^m$ be connected nicely curved submanifolds of a complete simply-connected manifold N of constant curvature. Let $\phi: M \rightarrow \bar{M}$ be an isometry. Suppose that for all $k \geq 1$ there are bundle isomorphisms $\tilde{\phi}^k: \text{Nor}^k M \rightarrow \text{Nor}^k \bar{M}$ covering ϕ which preserve inner products, second fundamental forms s^k , and connections D^k . Then there is an isometry A of N such that $\phi = A|_M$ and $\tilde{\phi}^k = A_*|_{\text{Nor}^k M}$.

PROOF. We obviously want to reduce this to Theorem 20. Notice first that since $\text{Osc}^t M_p = \text{Nor}^0 M_p \oplus \cdots \oplus \text{Nor}^{t-1} M_p$, and similarly for \bar{M} , the formal imbedding dimension $\#(M)$ must equal $\#(\bar{M})$. Taking into account Proposition 67, we see that there is no loss of generality in assuming that $\#(M) = \#(\bar{M}) = m$. Then the bundle isomorphisms $\tilde{\phi}^1, \dots, \tilde{\phi}^{l-1}$ combine to give a bundle isomorphism $\tilde{\phi}: \text{Nor} M \rightarrow \text{Nor} \bar{M}$. Clearly $\tilde{\phi}$ preserves inner products, second fundamental forms s^k , and connections D^k . In particular $\tilde{\phi}$ takes the second fundamental form $s (= s^1)$ to $\bar{s} (= \bar{s}^1)$. To prove that

$$(*) \quad \tilde{\phi}(D_X \xi) = \bar{D}_{\phi_* X}(\tilde{\phi}(\xi))$$

for all sections ξ of E , it suffices to consider separately sections ξ of $\text{Nor}^k M$. Then the Frenet equations give

$$D_X \xi = \perp(\nabla'_X \xi) = \begin{cases} D^1_X \xi + s^1(X, \xi) & k = 1 \\ -A^k_\xi(X) + D^k_X \xi + s^k(X, \xi) & k > 1, \end{cases}$$

with corresponding formulas for $\bar{D}_{\phi_* X} \tilde{\phi}(\xi)$. Since $\tilde{\phi}$ preserves D^k , as well as A^k and s^k (for they are determined by s^{k-1} , s^k and s^k), we see that equation (*) does indeed hold. ♦

Now we need certain equations satisfied by the connections D^k . We will state these in terms of vector fields on M and sections of $\text{Nor}^k M$. After the proof we will give another formulation, in terms of tangent vectors in M_p , and vectors in $\text{Nor}^k M_p$, which will make the result appear as a genuine generalization of the Codazzi-Mainardi equations for D .

72. THEOREM. Let $M \subset N$ be nicely curved. Then for all vector fields X, Y on M and sections ξ of $\text{Nor}^k M$ ($k \geq 0$) we have

The Generalized Codazzi-Mainardi Equations:

$$\begin{aligned} \mathsf{T}^{k+1} R'(X, Y)\xi &= [D^{k+1}_X(s^k(Y, \xi)) - s^k(\nabla_X Y, \xi) - s^k(Y, D^k_X \xi)] \\ &\quad - [D^{k+1}_Y(s^k(X, \xi)) - s^k(\nabla_Y X, \xi) - s^k(X, D^k_Y \xi)]. \end{aligned}$$

When N has constant curvature, the left side is zero.

PROOF. By the Frenet equations we have

$$\nabla'_Y \xi = -A^k_\xi(Y) + D^k_Y \xi + s^k(Y, \xi).$$

Since $A^k_\xi(Y)$ is a section of $\text{Nor}^{k-1} M$, Lemma 69 implies that

$$\mathsf{T}^{k+1} \nabla'_X \nabla'_Y \xi = \mathsf{T}^{k+1} \nabla'_X D^k_Y \xi + \mathsf{T}^{k+1} \nabla'_X s^k(Y, \xi).$$

Using (***) on page 170, and the definition of D^{k+1} , we thus have

$$(1) \quad \mathsf{T}^{k+1} \nabla'_X \nabla'_Y \xi = s^k(X, D^k_Y \xi) + D^{k+1}_X(s^k(Y, \xi)),$$

$$(1') \quad \mathsf{T}^{k+1} \nabla'_Y \nabla'_X \xi = s^k(Y, D^k_X \xi) + D^{k+1}_Y(s^k(X, \xi)).$$

We also have, by (***),

$$\mathsf{T}^{k+1} \nabla'_{[X, Y]} \xi = s^k([X, Y], \xi),$$

and thus

$$(2) \quad \mathsf{T}^{k+1} \nabla'_{[X,Y]} \xi = \mathbf{s}^k(\nabla_X Y, \xi) - \mathbf{s}^k(\nabla_Y X, \xi).$$

Substituting (1), (1'), (2) into the formula $R'(X, Y)\xi = \nabla'_X \nabla'_Y \xi - \nabla'_Y \nabla'_X \xi - \nabla'_{[X,Y]} \xi$, we obtain the desired result.

When N has constant curvature K_0 we have

$$R'(X, Y)\xi = K_0[\langle Y, \xi \rangle X - \langle X, \xi \rangle Y],$$

which is tangent to M . So $\mathsf{T}^{k+1} R'(X, Y)\xi = 0$. \blacklozenge

It is easily checked that in these generalized Codazzi-Mainardi equations, each of the expressions in brackets is linear in X , Y , and ξ over the C^∞ functions, and thus its value at p depends only on X_p, Y_p, ξ_p . To give this value explicitly, we note that we can consider \mathbf{s}^k as a section of the bundle $\text{Hom}(TM \times \text{Nor}^k M, \text{Nor}^{k+1} M)$. Using the connections ∇ , D^k , and D^{k+1} on TM , $\text{Nor}^k M$, and $\text{Nor}^{k+1} M$, we can define a natural connection $\tilde{\nabla}$ on this bundle (compare page 37 for the case $k = 0$). It is easily seen that Theorem 72 can be written

$$\begin{aligned} \mathsf{T}^{k+1} R'(X_p, Y_p)\xi_p &= (\tilde{\nabla}_{X_p} \mathbf{s}^k)(Y_p, \xi_p) - (\tilde{\nabla}_{Y_p} \mathbf{s}^k)(X_p, \xi_p) \\ X_p, Y_p &\in M_p, \text{ and } \xi_p \in \text{Nor}^k M_p. \end{aligned}$$

Now we can state the proper form of Lemma 65.

73. LEMMA (FUNDAMENTAL LEMMA OF RIEMANNIAN SUBMANIFOLD THEORY). Let $M \subset N$ be nicely curved. Then the set of normal connections D^k in $\text{Nor}^k M$ is the unique set of connections δ^k on $\text{Nor}^k M$ such that

$$(0) \quad \delta^0 = \nabla,$$

$$(1) \quad \delta^k \text{ is compatible with the metric in } \text{Nor}^k M:$$

$$X(\langle \xi, \eta \rangle) = \langle \delta_X^k \xi, \eta \rangle + \langle \xi, \delta_X^k \eta \rangle \quad \text{for sections } \xi, \eta \text{ of } \text{Nor}^k M,$$

$$(2) \quad \text{The } \delta^k \text{ satisfy the Codazzi-Mainardi equations:}$$

$$\begin{aligned} \mathsf{T}^{k+1} R'(X, Y)\xi &= [\delta^{k+1}_X (\mathbf{s}^k(Y, \xi)) - \mathbf{s}^k(\nabla_X Y, \xi) - \mathbf{s}^k(Y, \delta_X^k \xi)] \\ &\quad - [\delta^{k+1}_Y (\mathbf{s}^k(X, \xi)) - \mathbf{s}^k(\nabla_Y X, \xi) - \mathbf{s}^k(X, \delta_Y^k \xi)]. \end{aligned}$$

PROOF. We will show that if $\delta^k = D^k$, then $\delta^{k+1} = D^{k+1}$. Since $\delta^0 = \nabla = D^0$, this will prove the result. We begin by considering the expression

$$\begin{aligned} & \langle \delta^{k+1}_X \mathbf{s}^k(Y_1, \xi), \mathbf{s}^k(Z_1, \eta) \rangle - \langle \delta^{k+1}_{Y_1} \mathbf{s}^k(X, \xi), \mathbf{s}^k(Z_1, \eta) \rangle \\ & - \langle \delta^{k+1}_{Y_1} \mathbf{s}^k(Z_1, \eta), \mathbf{s}^k(X, \xi) \rangle + \langle \delta^{k+1}_{Z_1} \mathbf{s}^k(Y_1, \eta), \mathbf{s}^k(X, \xi) \rangle \\ & + \langle \delta^{k+1}_{Z_1} \mathbf{s}^k(X, \xi), \mathbf{s}^k(Y_1, \eta) \rangle - \langle \delta^{k+1}_X \mathbf{s}^k(Z_1, \xi), \mathbf{s}^k(Y_1, \eta) \rangle, \end{aligned}$$

where ξ, η are sections of $\text{Nor}^k M$. As in the proof of Lemma 65, we are led to the conclusion that we can write

$$\langle \delta^{k+1}_X \mathbf{s}^k(Y_1, \xi), \mathbf{s}^k(Z_1, \eta) \rangle - \langle \delta^{k+1}_X \mathbf{s}^k(Z_1, \xi), \mathbf{s}^k(Y_1, \eta) \rangle = \dots,$$

where \dots can be expressed in terms of $X, Y_1, Z_1, \xi, \eta, \delta^k = D^k$. In particular, if we choose $\xi = s^k(Y_2, \dots, Y_{k+2})$ and $\eta = s^k(Z_2, \dots, Z_{k+2})$, then we obtain

$$\begin{aligned} (3) \quad & \langle \delta^{k+1}_X s^{k+1}(Y_1, \dots, Y_{k+2}), s^{k+1}(Z_1, \dots, Z_{k+2}) \rangle \\ & - \langle \delta^{k+1}_X s^{k+1}(Z_1, Y_2, \dots, Y_{k+2}), s^{k+1}(Y_1, Z_2, \dots, Z_{k+2}) \rangle = \dots \end{aligned}$$

We will abbreviate the left side of this equation by

$$\{Y_1, \dots, Y_{k+2}; Z_1, \dots, Z_{k+2}\} - \{Z_1, Y_2, \dots, Y_{k+2}; Y_1, Z_2, \dots, Z_{k+2}\}.$$

Now consider the following expressions (the pattern becomes apparent by looking at the terms after the $-$ signs):

$$\begin{aligned} & \{Y_1, Y_2, Y_3, \dots, Y_{k+2}; Z_1, Z_2, Z_3, \dots, Z_{k+2}\} \\ & \quad - \{Z_1, Y_2, Y_3, \dots, Y_{k+2}; Y_1, Z_2, Z_3, \dots, Z_{k+2}\} \\ & \{Y_2, Z_1, Y_3, \dots, Y_{k+2}; Z_2, Y_1, Z_3, \dots, Z_{k+2}\} \\ & \quad - \{Z_2, Z_1, Y_3, \dots, Y_{k+2}; Y_2, Y_1, Z_3, \dots, Z_{k+2}\} \\ & \{Y_3, Z_2, Z_1, \dots, Y_{k+2}; Z_3, Y_2, Y_1, \dots, Z_{k+2}\} \\ & \quad - \{Z_3, Z_2, Z_1, \dots, Y_{k+2}; Y_3, Y_2, Y_1, \dots, Z_{k+2}\} \\ & \quad \vdots \\ & \{Y_{k+1}, Z_k, \dots, Z_1, Y_{k+2}; Z_{k+1}, Y_k, \dots, Y_1, Z_{k+2}\} \\ & \quad - \{Z_{k+1}, \dots, Z_1, Y_{k+2}; Y_{k+1}, \dots, Y_1, Z_{k+2}\} \\ & \{Y_{k+2}, Z_{k+1}, \dots, Z_1; Z_{k+2}, Y_{k+1}, \dots, Y_1\} \\ & \quad - \{Z_{k+2}, \dots, Z_1; Y_{k+2}, \dots, Y_1\}. \end{aligned}$$

Notice that each term after a $-$ sign is the same as the term on the next line, since s^{k+1} is symmetric. So adding all the equations (3) having the above expressions on the left we obtain

$$(4) \quad \langle \delta^{k+1}_X s^{k+1}(Y_1, \dots, Y_{k+2}), s^{k+1}(Z_1, \dots, Z_{k+2}) \rangle \\ - \langle \delta^{k+1}_X s^{k+1}(Z_1, \dots, Z_{k+2}), s^{k+1}(Y_1, \dots, Y_{k+2}) \rangle = \dots$$

But by (1) we also have

$$(5) \quad \langle \delta^{k+1}_X s^{k+1}(Y_1, \dots, Y_{k+2}), s^{k+1}(Z_1, \dots, Z_{k+2}) \rangle \\ + \langle \delta^{k+1}_X s^{k+1}(Z_1, \dots, Z_{k+2}), s^{k+1}(Y_1, \dots, Y_{k+2}) \rangle = \dots$$

So by adding (4) and (5) we obtain

$$(*) \quad \langle \delta^{k+1}_X s^{k+1}(Y_1, \dots, Y_{k+2}), s^{k+1}(Z_1, \dots, Z_{k+2}) \rangle = \dots$$

Since $\text{Nor}^{k+1} M_p$ is spanned by image s^{k+1} , this proves, as in Lemma 65, that δ^{k+1} is uniquely determined. \blacklozenge

Now for a manifold $M \subset N$ we define tensors \mathcal{F}_k by

$$\mathcal{F}_k(X_1, \dots, X_{k+1}, Y_1, \dots, Y_{k+1}) = \langle s^k(X_1, \dots, X_{k+1}), s^k(Y_1, \dots, Y_{k+1}) \rangle.$$

If $X_1, \dots, X_n \in M_p$ is a basis, then we can form the $n^{k+1} \times n^{k+1}$ matrix

$$(\mathcal{F}_k(X_{i_1}, \dots, X_{i_{k+1}}, X_{j_1}, \dots, X_{j_{k+1}})).$$

It is easy to see that this matrix is positive semi-definite and that its rank is just the dimension of ${}^* \text{Nor}^k M_p$.

74. THEOREM. Let $M, \bar{M} \subset N$ be connected nicely curved submanifolds of a complete simply-connected manifold N of constant curvature. Let $\phi: M \rightarrow \bar{M}$ be an isometry such that

$$\phi^* \bar{\mathcal{F}}_k = \mathcal{F}_k \quad \text{for all } k.$$

Then ϕ is the restriction of an isometry of N .

* For those who know about tensor products of vector spaces this can be expressed more simply: We can regard s^k as a linear map $s^k: M_p \otimes \dots \otimes M_p \rightarrow \text{Nor}^k M_p$, so \mathcal{F}_k is a bilinear map $\mathcal{F}_k: (M_p \otimes \dots \otimes M_p) \times (M_p \otimes \dots \otimes M_p) \rightarrow \mathbb{R}$. The matrix considered above is the matrix of this bilinear map with respect to the basis $\{X_{i_1} \otimes \dots \otimes X_{i_{k+1}}\}$ of $M_p \otimes \dots \otimes M_p$.

PROOF. The preceding remarks show that the dimension $\text{Nor}^k M$ must equal the dimension of $\text{Nor}^k \bar{M}$. Since s^k is onto $\text{Nor}^k M$, the procedure used in the proof of Proposition 66 allows us to construct bundle isomorphisms $\tilde{\phi}^k: \text{Nor}^k M \rightarrow \text{Nor}^k \bar{M}$ which preserve inner products and second fundamental forms s^k . Again arguing as in Proposition 66, but using Lemma 73 in place of Lemma 65, we see that the $\tilde{\phi}^k$ also preserve the connections D^k . So we can apply Theorem 71. ♦

We would also like to discuss when a given set of tensors $\{\mathcal{F}_k\}$ on a manifold M come from an imbedding of M in a complete manifold N of constant curvature. The Codazzi-Mainardi equations represent only one set of integrability conditions, and we still have to consider the other components of $\nabla'_X \nabla'_Y \xi - \nabla'_Y \nabla'_X \xi - \nabla'_{[X,Y]} \xi$. If ξ is a section of $\text{Nor}^k M$, then the only components we have to consider are $\mathsf{T}^{k-2}, \dots, \mathsf{T}^{k+2}$, where T^{k+1} is already taken care of by the Codazzi-Mainardi equations.

First consider T^{k+2} . From the Frenet equations

$$\nabla'_Y \xi = -A_\xi^k(Y) + D_Y^k \xi + s^k(Y, \xi)$$

we obtain

$$\mathsf{T}^{k+2} \nabla'_X \nabla'_Y \xi = \mathsf{T}^{k+2} \nabla'_X s^k(Y, \xi) = s^{k+1}(X, s^k(Y, \xi)) \quad \text{by (***)}.$$

Also

$$\begin{aligned} \mathsf{T}^{k+2} \nabla'_Y \nabla'_X \xi &= s^{k+1}(Y, s^k(X, \xi)) \\ \mathsf{T}^{k+2} \nabla'_{[X,Y]} \xi &= 0. \end{aligned}$$

So we have

$$\mathsf{T}^{k+2} R'(X, Y) \xi = s^{k+1}(X, s^k(Y, \xi)) - s^{k+1}(Y, s^k(X, \xi)).$$

In a space of constant curvature, the left side is 0. On the other hand, the right hand side is clearly always 0, since s^{k+2} is symmetric. Thus we do not obtain any new condition for imbedding in a manifold of constant curvature by looking at T^{k+2} .

Next consider T^{k-2} . The Frenet equations give us [recall the alternative notation $A^k(\xi; X)$ for $A_\xi^k(X)$]

$$\begin{aligned} \mathsf{T}^{k-2} \nabla'_X \nabla'_Y \xi &= -\mathsf{T}^{k-2} \nabla'_X A_\xi^k(Y) = A^{k-1}(A_\xi^k(Y); X) \\ \mathsf{T}^{k-2} \nabla'_Y \nabla'_X \xi &= -\mathsf{T}^{k-2} \nabla'_Y A_\xi^k(X) = A^{k-1}(A_\xi^k(X); Y) \\ \mathsf{T}^{k-2} \nabla'_{[X,Y]} \xi &= 0. \end{aligned}$$

So we obtain

$$\mathbb{T}^{k-2} R'(X, Y)\xi = A^{k-1}(A_\xi^k(Y); X) - A^{k-1}(A_\xi^k(X); Y).$$

In a space of constant curvature the left side is 0 (this is clear for $k > 2$, since $R'(X, Y)\xi$ is tangent to M ; it is true even for $k = 2$, since $R'(X, Y)\xi = K_0[\langle Y, \xi \rangle X - \langle X, \xi \rangle Y]$, and $\langle X, \xi \rangle = \langle Y, \xi \rangle = 0$). On the other hand, for any section η of Nor^{k-2} we have

$$\begin{aligned} \langle A^{k-1}(A_\xi^k(Y); X), \eta \rangle &= \langle A_\xi^k(Y), \mathbf{s}^{k-1}(X, \eta) \rangle \\ &= \langle \xi, \mathbf{s}^k(Y, \mathbf{s}^{k-1}(X, \eta)) \rangle, \end{aligned}$$

so we see that the right side of our equation is always 0. So, once again, we obtain no new conditions for imbedding M in a manifold of constant curvature.

Now consider \mathbb{T}^{k-1} . We have

$$\begin{aligned} \mathbb{T}^{k-1} \nabla'_X \nabla'_Y \xi &= -\mathbb{T}^{k-1} \nabla'_X A_\xi^k(Y) + \mathbb{T}^{k-1} \nabla'_X D_Y^k \xi \\ &= -D_X^{k-1} A_\xi^k(Y) - A^k(D_Y^k \xi; X) \\ \mathbb{T}^{k-1} \nabla'_Y \nabla'_X \xi &= -D_Y^{k-1} A_\xi^k(X) - A^k(D_X^k \xi; Y) \\ \mathbb{T}^{k-1} \nabla'_{[X, Y]} \xi &= -A_\xi^k([X, Y]) = -A_\xi^k(\nabla_X Y) + A_\xi^k(\nabla_Y X). \end{aligned}$$

Thus we obtain

$$\begin{aligned} -\mathbb{T}^{k-1} R'(X, Y)\xi &= [D_X^{k-1} A_\xi^k(Y) - A^k(D_X^k \xi; Y) - A_\xi^k(\nabla_X Y)] \\ &\quad - [D_Y^{k-1} A_\xi^k(X) - A^k(D_Y^k \xi; X) - A_\xi^k(\nabla_Y X)]. \end{aligned}$$

Taking the inner product with a section η of $\text{Nor}^{k-1} M$, we obtain the equivalent equation

$$\begin{aligned} \text{(a)} \quad -\langle R'(X, Y)\xi, \eta \rangle &= [\langle D_X^{k-1} A_\xi^k(Y), \eta \rangle - \langle D_X^k \xi, \mathbf{s}^{k-1}(Y, \eta) \rangle \\ &\quad - \langle \xi, \mathbf{s}^{k-1}(\nabla_X Y, \eta) \rangle] \\ &\quad - [\langle D_Y^{k-1} A_\xi^k(X), \eta \rangle - \langle D_Y^k \xi, \mathbf{s}^{k-1}(X, \eta) \rangle \\ &\quad - \langle \xi, \mathbf{s}^{k-1}(\nabla_Y X, \eta) \rangle]. \end{aligned}$$

But we also have

$$\begin{aligned} \langle A_\xi^k(Y), \eta \rangle &= \langle \xi, \mathbf{s}^{k-1}(Y, \eta) \rangle \\ \implies X(\langle A_\xi^k(Y), \eta \rangle) &= X(\langle \xi, \mathbf{s}^{k-1}(Y, \eta) \rangle) \end{aligned}$$

$$\begin{aligned}
&\implies \langle D^{k-1}_X A^k_\xi(Y), \eta \rangle + \langle A^k_\xi(Y), D^{k-1}_X \eta \rangle \\
&\quad = \langle D^k_X \xi, s^{k-1}(Y, \eta) \rangle + \langle \xi, D^k_X s^{k-1}(Y, \eta) \rangle \\
&\implies \langle D^{k-1}_X A^k_\xi(Y), \eta \rangle - \langle D^k_X \xi, s^{k-1}(Y, \eta) \rangle \\
&\quad = \langle \xi, D^k_X s^{k-1}(Y, \eta) \rangle - \langle \xi, s^{k-1}(Y, D^{k-1}_X \eta) \rangle.
\end{aligned}$$

Therefore the right side of (a) can be written

$$\begin{aligned}
&[\langle \xi, D^k_X s^{k-1}(Y, \eta) \rangle - \langle \xi, s^{k-1}(Y, D^{k-1}_X \eta) \rangle - \langle \xi, s^{k-1}(\nabla_X Y, \eta) \rangle] \\
&\quad - [\langle \xi, D^{k-1}_Y s^{k-1}(X, \eta) \rangle - \langle \xi, s^{k-1}(X, D^{k-1}_Y \eta) \rangle - \langle \xi, s^{k-1}(\nabla_Y X, \eta) \rangle] \\
&= \langle R'(X, Y)\eta, \xi \rangle, \quad \text{by the Codazzi-Mainardi equations.}
\end{aligned}$$

So equation (a) follows from the Codazzi-Mainardi equations; we obtain no new conditions by looking at T^{k-1} .

Finally, we have to look at T^k . We have

$$\begin{aligned}
T^k \nabla'_X \nabla'_Y \xi &= -T^k \nabla'_X A^k_\xi(Y) + T^k \nabla'_X D^k_Y \xi + T^k \nabla'_X s^k(Y, \xi) \\
&= -s^{k-1}(X, A^k_\xi(Y)) + D^k_X D^k_Y \xi - A^{k+1}(s^k(Y, \xi); X) \\
T^k \nabla'_Y \nabla'_X \xi &= -s^{k-1}(Y, A^k_\xi(X)) + D^k_Y D^k_X \xi - A^{k+1}(s^k(X, \xi); Y) \\
T^k \nabla'_{[X, Y]} \xi &= D^k_{[X, Y]} \xi.
\end{aligned}$$

So we obtain

$$\begin{aligned}
(b) \quad T^k R'(X, Y)\xi &= D^k_X D^k_Y \xi - D^k_Y D^k_X \xi - D^k_{[X, Y]} \xi \\
&\quad + s^{k-1}(X, A^k_\xi(Y)) - s^{k-1}(Y, A^k_\xi(X)) \\
&\quad + A^{k+1}(s^k(X, \xi); Y) - A^{k+1}(s^k(Y, \xi); X).
\end{aligned}$$

When $k = 0$, the terms involving s^{k-1} do not appear. In this case, if we take $\xi = Z$ to be a section of $\text{Nor}^0 M = TM$ we obtain

$$\begin{aligned}
R'(X, Y)Z &= R(X, Y)Z + A^1(s(X, Z); Y) - A^1(s(Y, Z); X) \\
&\quad \Downarrow \\
\langle R'(X, Y)Z, W \rangle &= \langle R(X, Y)Z, W \rangle + \langle s(X, Z), s(Y, W) \rangle - \langle s(X, W), s(Y, Z) \rangle
\end{aligned}$$

i.e., Gauss' equation. But for $k > 0$ we obtain an unsavory hybrid between Gauss' equation and the Ricci equations. We can obtain a nicer looking set of

equations by considering the bundles $\text{Osc}^k M$. There is a projection $\mathbb{T}^{[k]}: N_p \rightarrow \text{Osc}^k M_p$, defined by means of the orthogonal complement of $\text{Osc}^k M_p$ in N_p , and we can thus define a connection $D^{[k]}$ on $\text{Osc}^k M$ by

$$D^{[k]}_X \xi = \mathbb{T}^{[k]} \nabla'_X \xi \quad \xi \text{ a section of } \text{Osc}^k M.$$

This connection has a curvature tensor $R^{[k]}$ defined by

$$R^{[k]}(X, Y)\xi = D^{[k]}_X D^{[k]}_Y \xi - D^{[k]}_Y D^{[k]}_X \xi - D^{[k]}_{[X, Y]}\xi.$$

75. PROPOSITION. Let $M \subset N$ be nicely curved. Then for all vectors $X, Y \in M_p$ and $\xi \in \text{Osc}^k M_p$ we have the

Generalized Gauss Equation:

$$\begin{aligned} \mathbb{T}^{[k]} R'(X, Y)\xi = \\ R^{[k]}(X, Y)\xi + A^k(\mathbf{s}^{k-1}(X, \mathbb{T}^{k-1}\xi); Y) - A^k(\mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1}\xi); X). \end{aligned}$$

So for $\xi, \eta \in \text{Osc}^k M_p$ we have

$$\begin{aligned} \langle R'(X, Y)\xi, \eta \rangle = \langle R^{[k]}(X, Y)\xi, \eta \rangle + \langle \mathbf{s}^{k-1}(X, \mathbb{T}^{k-1}\xi), \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1}\eta) \rangle \\ - \langle \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1}\xi), \mathbf{s}^{k-1}(X, \mathbb{T}^{k-1}\eta) \rangle. \end{aligned}$$

PROOF. We have

$$\begin{aligned} \nabla'_Y \xi &= D^{[k]}_Y \xi + \mathbb{T}^k \nabla'_Y \xi \\ &= D^{[k]}_Y \xi + \mathbb{T}^k (\nabla'_Y \mathbb{T}^{k-1} \xi) \\ &= D^{[k]}_Y \xi + \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1} \xi). \end{aligned}$$

Therefore

$$\begin{aligned} \nabla'_X \nabla'_Y \xi &= D^{[k]}_X D^{[k]}_Y \xi + \mathbf{s}^{k-1}(X, \mathbb{T}^{k-1} \cdot D^{[k]}_Y \xi) \\ &\quad + D^{[k]}_X \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1} \xi) + 0. \end{aligned}$$

So

$$\begin{aligned} (1) \quad \mathbb{T}^{[k]} \nabla'_X \nabla'_Y \xi &= D^{[k]}_X D^{[k]}_Y \xi + D^{[k]}_X \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1} \xi) \\ &= D^{[k]}_X D^{[k]}_Y \xi + \mathbb{T}^{k-1} \nabla'_X \mathbf{s}^{k-1}(Y, \mathbb{T}^{k-1} \xi) \\ &= D^{[k]}_X D^{[k]}_Y \xi - A^k(\mathbf{s}^{k-1}(Y, \mathbf{s}^{k-1} \xi); X). \end{aligned}$$

Also

$$(2) \quad \mathbb{T}^{[k]} \nabla'_Y \nabla'_X \xi = D^{[k]}_Y D^{[k]}_X \xi - A^k(\mathbf{s}^{k-1}(X, \mathbb{T}^{k-1} \xi); Y)$$

$$(3) \quad \mathbb{T}^{[k]} \nabla'_{[X, Y]} \xi = D^{[k]}_{[X, Y]} \xi.$$

Equations (1)–(3) give the result. ♦

Although we derived Gauss' equation from scratch, it is important to note that it is formally equivalent to equation (b) on page 179, in the following sense. For a section ξ of $\text{Osc}^k M$ we could define $D^{[k]}_X \xi$ as

$$\begin{aligned} D^{[k]}_X \xi = & [D^0_X \mathbb{T}^0 \xi + s^0(X, \mathbb{T}^0 \xi)] \\ & + [-A^1(\mathbb{T}^1 \xi; X) + D^1_X \mathbb{T}^1 \xi + s^1(X, \xi)] \\ & \vdots \\ & + [-A^{k-2}(\mathbb{T}^{k-2} \xi; X) + D^{k-2}_X \mathbb{T}^{k-2} \xi + s^{k-2}(X, \xi)] \\ & + [-A^{k-1}(\mathbb{T}^{k-1} \xi; X) + D^{k-1}_X \mathbb{T}^{k-1} \xi]. \end{aligned}$$

Then the equations of Proposition 75, together with the Codazzi-Mainardi equations, imply equations (b) on page 179; the verification of this claim is left to the reader. So the Codazzi-Mainardi equations and Gauss' equation are the full set of integrability conditions for the Frenet equations. But we still have a lot of work to do before we can decide when a set of tensors $\{\mathcal{F}_k\}$ on M come from an imbedding of M in a space of constant curvature.

First we claim that if ℓ has its usual significance, then

$$\begin{aligned} R^{[\ell]}(X, Y)\xi &= \mathbb{T}^{[\ell]} R'(X, Y)\xi \\ &= 0, \text{ when } N \text{ has constant curvature.} \end{aligned}$$

This follows immediately from Proposition 67, which shows that $D^{[\ell]} = \nabla'$ on Osc^ℓ . We could also note that $R^{[\ell]} = R^{[\ell+1]}$, and that the terms $A^{\ell+1}$ which then arise in Gauss' equation are 0, since they lie in $\text{Nor}^\ell M_p$.

Now we have to establish certain important identities for the curvature tensors $R^{[k]}$, analogous to those for $R = R^{[1]}$. Recall that we have

- (1) $R(X, Y)Z + R(Y, X)Z = 0$
- (2) $\langle R(X, Y)Z, W \rangle + \langle R(X, Y)W, Z \rangle = 0$
- (3) $\mathfrak{S}\{R(X, Y)Z\} = R(X, Y)Z + R(Y, Z)X + R(Z, X)Y = 0$
- (4) $\langle R(X, Y)Z, W \rangle + \langle R(Z, W)X, Y \rangle = 0.$

When we are dealing with a submanifold M of another Riemannian manifold N , these identities follow immediately from Gauss' equation

$$\langle R'(X, Y)Z, W \rangle = \langle R(X, Y)Z, W \rangle + \langle s(X, Z), s(Y, W) \rangle - \langle s(Y, Z), s(X, W) \rangle,$$

and the corresponding identities for R' . Similarly, we have

76. PROPOSITION. Let $M \subset N$ be nicely curved. Then

$$(1) \quad R^{[k]}(X, Y)\xi + R^{[k]}(Y, X)\xi = 0$$

$$(2) \quad \langle R^{[k]}(X, Y)\xi, \eta \rangle + \langle R^{[k]}(X, Y)\eta, \xi \rangle = 0$$

$$(3) \quad \mathfrak{S}\{R^{[k]}(X, Y) \cdot s^{k-2}(Z, \zeta)\} = 0 \quad \zeta \in \text{Nor}^{k-2} M_p$$

$$(3') \quad \mathfrak{S}'\{\langle R^{[k]}(X, Y) \cdot s^{k-1}(Z_1, \dots, Z_k), s^{k-1}(W_1, \dots, W_k) \rangle\} = 0$$

where \mathfrak{S}' indicates a cyclic sum over $(Y, Z_1, \dots, Z_k, W_1, \dots, W_k)$

$$(4) \quad 0 = \langle R^{[k]}(X_1, Y_1) \cdot s^{k-1}(X_2, \dots, X_{k+1}), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ + \langle R^{[k]}(X_2, Y_2) \cdot s^{k-1}(Y_1, X_3, \dots, X_{k+1}), s^{k-1}(X_1, Y_3, \dots, Y_{k+1}) \rangle \\ + \langle R^{[k]}(X_3, Y_3) \cdot s^{k-1}(Y_1, Y_2, X_4, \dots, X_{k+1}), s^{k-1}(X_1, X_2, Y_4, \dots, Y_{k+1}) \rangle \\ \vdots \\ + \langle R^{[k]}(X_{k+1}, Y_{k+1}) \cdot s^{k-1}(Y_1, \dots, Y_k), s^{k-1}(X_1, \dots, X_k) \rangle.$$

Moreover, these identities follow formally from Gauss' equation for $R^{[k]}$ (and the properties of the curvature tensor R' for the ambient manifold).

PROOF. An easy computation. ♦

More important for us will be the (second) Bianchi identity

$$\mathfrak{S}\{(\nabla_Z R)(X, Y, W)\} = 0$$

[where we write R as $(X, Y, W) \mapsto R(X, Y, W)$].

Although we have derived this identity for the curvature tensor of a (symmetric) connection on the tangent bundle, it is actually more general:

77. PROPOSITION. Let ∇ be a connection on TM , with torsion tensor T , and let D be a connection on a bundle $\varpi: E \rightarrow M$ with curvature tensor $R = R_D$. Let $\tilde{\nabla}$ be the natural connection on the bundle $\text{Hom}(TM \times TM \times E, E)$ determined by the connections ∇ on TM and D on E . Then

$$\mathfrak{S}\{(\tilde{\nabla}_Z R)(X, Y, \xi)\} + \mathfrak{S}\{R(T(X, Y), Z)\xi\} = 0.$$

In particular, if $T = 0$, then

$$\mathfrak{S}\{(\tilde{\nabla}_Z R)(X, Y, \xi)\}.$$

PROOF. Exactly the same as the proof on pp. II.244–245, replacing W by ξ throughout. ♦

Note that when E is TM , the connection $\tilde{\nabla}$ is just denoted by ∇ , in conformity with previous usage.

78. COROLLARY. Let $M \subset N$ be nicely curved, and let $\tilde{\nabla}$ be the natural connection on $\text{Hom}(TM \times TM \times \text{Osc}^k M, \text{Osc}^k M)$ determined by the connections ∇ on TM and $D^{[k]}$ on $\text{Osc}^k M$. Then

$$(1) \quad \mathfrak{S}\{(\tilde{\nabla}_Z R^{[k]})(X, Y, \xi)\} = 0.$$

In addition,

$$(2) \quad \mathfrak{S}''\{(\tilde{\nabla}_{X_1} R^{[k]})(X, Y_1, s^{k-1}(X_2, \dots, X_{k+1}), s^{k-1}(Y_2, \dots, Y_{k+1}))\} = 0$$

where \mathfrak{S}'' indicates a cyclic sum over $(X_1, \dots, X_{k+1}, Y_1, \dots, Y_{k+1})$.

Moreover, equation (2) follows formally from (1), Gauss' equation for $R^{[k]}$, and the fact that the connection $D^{[k]}$ on $\text{Osc}^k M$ is compatible with the metric (and the properties of the curvature tensor R' for the ambient manifold).

PROOF. To obtain equation (2), we apply X to both sides of equation (4) in Proposition 76. We have, for example,

$$\begin{aligned} & X(\langle R^{[k]}(X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1})), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle) \\ &= \langle D^{[k]}_X(R^{[k]}(X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1}))), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ &\quad + \langle R^{[k]}(X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1})), D^{[k]}_X s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle, \end{aligned}$$

which by Corollary II.6.5 is

$$\begin{aligned} &= \langle (\tilde{\nabla}_X R^{[k]})(X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1})), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ &\quad + \langle R^{[k]}(\nabla_X X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1})), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ &\quad + \langle R^{[k]}(X_1, \nabla_X Y_1, s^{k-1}(X_2, \dots, X_{k+1})), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ &\quad + \langle R^{[k]}(X_1, Y_1, D^{[k]}_X s^{k-1}(X_2, \dots, X_{k+1})), s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle \\ &\quad + \langle R^{[k]}(X_1, Y_1, s^{k-1}(X_2, \dots, X_{k+1})), D^{[k]}_X s^{k-1}(Y_2, \dots, Y_{k+1}) \rangle. \end{aligned}$$

Using (1) we can replace the term involving $(\tilde{\nabla}_X R^{[k]})$ by two terms, involving $\tilde{\nabla}_{X_1} R^{[k]}$ and $\tilde{\nabla}_{Y_1} R^{[k]}$. After performing this substitution, and summing all the terms thus arising from equation (4) of Proposition 76, everything cancels out except for the terms which constitute equation (2). ♦

Corollary 78 will play an especially important role in our theory. To begin with, consider the case $R = R^{[1]}$, which depends only on the connection ∇ on TM . In the Remark after Lemma 65, we pointed out that for any bundle $\varpi: E \rightarrow M$ with a metric $\langle \cdot, \cdot \rangle$ and a symmetric section s of $\text{Hom}(TM \times TM, E)$, we can consider the ‘‘Codazzi-Mainardi equations’’ for a connection δ on E . The proof of Lemma 65 shows that if δ is to be compatible with the metric $\langle \cdot, \cdot \rangle$ and also satisfy this equation, then $\langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle$ is completely determined, by equation (*) in the proof. However, if we are given a δ which does satisfy (*), it is by no means clear that δ is compatible with the metric and satisfies the Codazzi-Mainardi equations. To see what is happening here, we need to examine the formulas much more closely. Returning to the proof of Lemma 65 one can see that when explicitly written out, equation (3) in the proof reads

$$\begin{aligned}
& \langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle - \langle \delta_X s(Z_1, Y_2), s(Y_1, Z_2) \rangle \\
&= \langle s(\nabla_{Y_1} X, Y_2) - s(\nabla_X Y_1, Y_2) + s(X, \nabla_{Y_1} Y_2) - s(Y_1, \nabla_X Y_2), s(Z_1, Z_2) \rangle \\
&- \langle s(\nabla_{Z_1} Y_1, Y_2) - s(\nabla_{Y_1} Z_1, Y_2) + s(Y_1, \nabla_{Z_1} Z_2) - s(Z_1, \nabla_{Y_1} Z_2), s(X, Y_2) \rangle \\
&+ \langle s(\nabla_X Z_1, Y_2) - s(\nabla_{Z_1} X, Y_2) + s(Z_1, \nabla_X Y_2) - s(X, \nabla_{Z_1} Y_2), s(Y_1, Z_2) \rangle \\
&+ Y_1(\langle s(X, Y_2), s(Z_1, Z_2) \rangle) - Z_1(\langle s(X, Y_2), s(Y_1, Z_2) \rangle) \\
&= \mathcal{E}(X, Y_1, Y_2, Z_1, Z_2), \text{ say.}
\end{aligned}$$

Following the proof a little further along, we arrive at the explicit formula

$$\begin{aligned}
2\langle \delta_X s(Y_1, Y_2), s(Z_1, Z_2) \rangle &= \mathcal{E}(X, Y_1, Y_2, Z_1, Z_2) + \mathcal{E}(X, Y_2, Z_1, Z_2, Y_1) \\
&+ X(\langle s(Y_1, Y_2), s(Z_1, Z_2) \rangle).
\end{aligned}$$

Now we can form

$$\begin{aligned}
& 2\langle \delta_U s(V, X), s(Y, Z) \rangle - 2\langle \delta_V s(U, X), s(Y, Z) \rangle \\
&= \mathcal{E}(U, V, X, Y, Z) \\
&\quad + \mathcal{E}(U, X, Y, Z, V) \\
&\quad - \mathcal{E}(V, U, X, Y, Z) \\
&\quad - \mathcal{E}(V, X, Y, Z, U) \\
&\quad + U(\langle s(V, X), s(Y, Z) \rangle) - V(\langle s(U, X), s(Y, Z) \rangle) \\
&= V(\langle s(U, X), s(Y, Z) \rangle) - Y(\langle s(U, X), s(V, Z) \rangle) + \cdots \\
&\quad + X(\langle s(U, Y), s(V, Z) \rangle) - Z(\langle s(U, Y), s(X, V) \rangle) + \cdots \\
&\quad - U(\langle s(V, X), s(Y, Z) \rangle) - Y(\langle s(V, X), s(U, Z) \rangle) + \cdots \\
&\quad - X(\langle s(V, Y), s(U, Z) \rangle) + Z(\langle s(V, Y), s(X, U) \rangle) + \cdots \\
&\quad + U(\langle s(V, X), s(Y, Z) \rangle) - V(\langle s(U, X), s(Y, Z) \rangle)
\end{aligned}$$

$$\begin{aligned}
&= \mathfrak{Z}\{Z(\langle s(X, U), s(V, Y) \rangle - \langle s(X, V), s(Y, U) \rangle)\} + \cdots \\
&\quad \text{where } \mathfrak{Z} \text{ indicates a cyclic sum over } (X, Y, Z) \\
&= \mathfrak{Z}\{Z(\langle R(X, Y)V, U \rangle)\} + \cdots \\
&= \mathfrak{Z}\{\langle \nabla_Z(R(X, Y)V), U \rangle\} + \cdots \\
&= \mathfrak{Z}\{\langle (\nabla_Z R)(X, Y, V), U \rangle\} + \cdots .
\end{aligned}$$

We have not troubled ourselves to write down all the \cdots terms, but, as you may suspect, when we apply Corollary 76(2) [for $k = 1$] we find that this equation comes down to precisely the Codazzi-Mainardi equations! In deriving this, we use only Gauss' equation for R , and the fact that ∇ is compatible with the metric (and properties of R' for the ambient manifold).

Similarly, we may form

$$\begin{aligned}
&2\langle \delta_X s(X_1, X_2), s(Y_1, Y_2) \rangle + 2\langle s(X_1, X_2), \delta_X s(Y_1, Y_2) \rangle \\
&= \mathcal{E}(X, X_1, X_2, Y_1, Y_2) \\
&\quad + \mathcal{E}(X, X_2, Y_1, Y_2, X_1) \\
&\quad + \mathcal{E}(X, Y_1, Y_2, X_1, X_2) \\
&\quad + \mathcal{E}(X, Y_2, X_1, X_2, Y_1) \\
&\quad + 2X(\langle s(X_1, X_2), s(Y_1, Y_2) \rangle) \\
&= X_1(\langle s(X, X_2), s(Y_1, Y_2) \rangle) - Y_1(\langle s(X, X_2), s(X_1, Y_2) \rangle) + \cdots \\
&\quad + X_2(\langle s(X, Y_1), s(Y_2, X_1) \rangle) - Y_2(\langle s(X, Y_1), s(X_2, X_1) \rangle) + \cdots \\
&\quad + Y_1(\langle s(X, Y_2), s(X_1, X_2) \rangle) - X_1(\langle s(X, Y_2), s(Y_1, X_2) \rangle) + \cdots \\
&\quad + Y_2(\langle s(X, X_1), s(X_2, Y_1) \rangle) - X_2(\langle s(X, X_1), s(Y_2, Y_1) \rangle) + \cdots \\
&\quad + 2X(\langle s(X_1, X_2), s(Y_1, Y_2) \rangle) \\
&= \mathfrak{Z}''\{X_1(\langle s(X, X_2), s(Y_1, Y_2) \rangle) - \langle s(X, Y_2), s(Y_1, X_2) \rangle\} + \cdots \\
&\quad \text{where } \mathfrak{Z}'' \text{ indicates a cyclic sum over } (X_1, X_2, Y_1, Y_2) \\
&= -\mathfrak{Z}''\{X_1(\langle R(X, Y_1)X_2, Y_2 \rangle)\} + \cdots \\
&= -\mathfrak{Z}''\{\langle \nabla_{X_1}(R(X, Y_1)X_2), Y_2 \rangle\} + \cdots \\
&= -\mathfrak{Z}''\{\langle (\nabla_{X_1} R)(X, Y_1, X_2), Y_2 \rangle\} + \cdots .
\end{aligned}$$

When we apply Corollary 78(2), it turns out that everything on the right side of this equation cancels, except the term $2X(\langle s(X_1, X_2), s(Y_1, Y_2) \rangle)$. So we see that δ is compatible with the metric!

More generally, we have

79. PROPOSITION. The fact that D^{k+1} satisfies the Codazzi-Mainardi equations and is compatible with the metric follows formally from equation (*) in

the proof of Lemma 73, Gauss' equation for $R^{[k+1]}$, and the fact that D^k is compatible with the metric (and the properties of R' for the ambient manifold).

PROOF. An abominable calculation. ♦

We are finally ready to consider the general imbedding question. The situation is rather complicated, and we will merely outline the results, without going into details. We are given a simply-connected manifold M^n and tensors $\mathcal{F}_0, \dots, \mathcal{F}_{\ell-1}$ on M , the tensor \mathcal{F}_k being covariant of order $2(k+1)$ and symmetric in the first $k+1$ arguments, in the last $k+1$ arguments, and under interchange of the first $k+1$ arguments with the last $k+1$ arguments. We assume that \mathcal{F}_0 is positive definite, and thus a Riemannian metric on M ; we will also denote \mathcal{F}_0 by $\langle \cdot, \cdot \rangle$. For $k \geq 1$ we assume that \mathcal{F}_k is positive semi-definite of constant rank $r_k > 0$. Set $m = n + r_1 + \dots + r_{\ell-1}$. We want to know when these tensors come from an immersion of M into a given complete m -dimensional manifold N of constant curvature K_0 . As usual, we can reduce this to a local problem, so we assume that M is diffeomorphic to \mathbb{R}^n , and we choose a basis X_1, \dots, X_n for the vector fields on M . For $1 \leq k \leq \ell - 1$ we take as our “ k^{th} normal bundle” $E^k = M \times \mathbb{R}^{r_k}$. Similarly, for our “ k^{th} osculating bundle” O^k we take the trivial bundle whose fibre over p is $O_p^k = M_p \oplus E_p^1 \oplus \dots \oplus E_p^{k-1}$. Each E^k has r_k natural sections $p \mapsto (p, (0, \dots, 0, 1, 0, \dots, 0))$, and we give E^k the Riemannian metric which makes these orthonormal; these metrics will all be denoted by $\langle \cdot, \cdot \rangle$. We now define $s^k: TM \times \dots \times TM \rightarrow E^k$ rather arbitrarily. By hypothesis, the $n^{k+1} \times n^{k+1}$ matrix

$$(\mathcal{F}_k(X_{i_1}, \dots, X_{i_{k+1}}, X_{j_1}, \dots, X_{j_{k+1}}))$$

has rank r_k at each point. Making M smaller if necessary, we can assume that there is a set \mathcal{S} of exactly r_k $(k+1)$ -tuples $(\alpha_1, \dots, \alpha_{k+1})$ such that the corresponding r_k rows of this matrix are everywhere linearly independent. Then for $(\alpha_1, \dots, \alpha_{k+1}) \in \mathcal{S}$ we define $s^k(X_{\alpha_1}, \dots, X_{\alpha_{k+1}})$ to be one of the r_k natural sections of E^k (choosing an arbitrary correspondence between the elements of \mathcal{S} and the r_k natural sections of E^k). There is now a unique way to define $s^k(X_{i_1}, \dots, X_{i_{k+1}})$ in general so that

$$\langle s^k(X_{i_1}, \dots, X_{i_{k+1}}), s^k(X_{j_1}, \dots, X_{j_{k+1}}) \rangle = \mathcal{F}_k(X_{i_1}, \dots, X_{i_{k+1}}, X_{j_1}, \dots, X_{j_{k+1}}).$$

Now we would like to define maps

$$s^k: TM \times E^k \rightarrow E^{k+1}$$

such that

$$s^k(X_i, s^k(X_{i_1}, \dots, X_{i_{k+1}})) = s^{k+1}(X_i, X_{i_1}, \dots, X_{i_{k+1}}).$$

But in this abstract set-up there is no way to prove that this map is well-defined. Instead we have to assume

(I) For each i and j , the $n^{k+1} \times 2n^{k+1}$ matrix

$$(\mathcal{F}_k(X_{i_1}, \dots, X_{i_{k+1}}, X_{j_1}, \dots, X_{j_{k+1}}), \mathcal{F}_{k+1}(X_i, X_{i_1}, \dots, X_{i_{k+1}}, X_j, X_{j_1}, \dots, X_{j_{k+1}}))$$

is of rank r_k . [The $(k+1)$ -tuple (i_1, \dots, i_{k+1}) determines a row of this matrix, and the $(k+1)$ -tuple (j_1, \dots, j_{k+1}) determines 2 different columns.]

With this assumption we can define s^k . We can thus also define the maps A_ξ^k for ξ an element of E^k .

Now we want to define connections D^k on the E^k . Consider first D^1 . The proof of Lemma 73 tells us that we have to define D^1 so that

$$(a_1) \quad \langle D^1_{X_i} s^1(X_{i_1}, X_{i_2}), s^1(X_{j_1}, X_{j_2}) \rangle = E_1(X_i, X_{i_1}, X_{i_2}, X_{j_1}, X_{j_2}),$$

where E_1 is some explicit expression we could work out. In order to know that we can define D^1 so that this formula holds, we must assume

(II₁) For each i, i_1, i_2 , the $n^2 \times 2n^2$ matrix

$$(\mathcal{F}_1(X_{h_1}, X_{h_2}, X_{j_1}, X_{j_2}), E_1(X_i, X_{i_1}, X_{i_2}, X_{j_1}, X_{j_2}))$$

is of rank r_2 . [The pair (h_1, h_2) determines a row, and the pair (j_1, j_2) determines 2 columns.]

With this assumption we can define D^1 so that equation (a₁) holds.

Of course, we already have the connection $D^0 = \nabla$ on TM determined by the metric $\mathcal{F}_0 = \langle \cdot, \cdot \rangle$, and we want to assume that its curvature tensor $R = R^{[1]}$ satisfies

$$\begin{aligned} \langle R'(X, Y)Z, W \rangle \\ = \langle R(X, Y)Z, W \rangle + \langle s^1(X, Z), s^1(Y, W) \rangle - \langle s^1(X, W), s^1(Y, Z) \rangle, \end{aligned}$$

i.e.,

$$\begin{aligned} (III_1) \quad K_0 \cdot [\langle X, W \rangle \cdot \langle Y, Z \rangle - \langle X, Z \rangle \cdot \langle Y, W \rangle] \\ = \langle R(X, Y)Z, W \rangle + \mathcal{F}_1(X, Z, Y, W) - \mathcal{F}_1(X, W, Y, Z). \end{aligned}$$

Proposition 79 then shows that D^1 satisfies the Codazzi-Mainardi equations and is compatible with the metric in E^1 . We can now define $D^{[2]}$ on O^2 by the formula on page 181, and it therefore makes sense to assume the generalized Gauss equation for $R^{[2]}$. Actually, it suffices to assume the special case

$$0 = \langle R^{[2]}(X, Y)s^1(X_1, X_2), s^1(Y_1, Y_2) \rangle \\ + \langle s^2(X, X_1, X_2), s^2(Y, Y_1, Y_2) \rangle - \langle s^2(Y, X_1, X_2), s^2(Y, Y_1, Y_2) \rangle,$$

i.e.,

$$(III_2) \quad 0 = \langle R^{[2]}(X, Y)s^1(X_1, X_2), s^1(Y_1, Y_2) \rangle \\ + \mathcal{F}_2(X, X_1, X_2, Y, Y_1, Y_2) - \mathcal{F}_2(Y, X_1, X_2, Y, Y_1, Y_2).$$

Now we want to define D^2 so that

$$(a_2) \quad \langle D^2_{X_i} s^2(X_{i_1}, X_{i_2}, X_{i_3}), s^2(X_{j_1}, X_{j_2}, X_{j_3}) \rangle \\ = E_2(X_i, X_{i_1}, X_{i_2}, X_{i_3}, X_{j_1}, X_{j_2}, X_{j_3}),$$

where E_2 is an explicit expression we could work out (it involves D^1 , but we already have an expression for D^1). In order to know that we can define D^2 so that this formula holds, we must assume

(II₂) For each i, i_1, i_2, i_3 , the $n^3 \times 2n^3$ matrix

$$(\mathcal{F}_2(X_{h_1}, X_{h_2}, X_{h_3}, X_{j_1}, X_{j_2}, X_{j_3}), E_2(X_i, X_{i_1}, X_{i_2}, X_{i_3}, X_{j_1}, X_{j_2}, X_{j_3}))$$

is of rank r_3 .

With this assumption we can define D^2 so that (a₂) holds. Then Proposition 79 shows that D^2 satisfies the Codazzi-Mainardi equations and is compatible with the metric in E^2 . We can now define $D^{[3]}$ on O^3 and it makes sense to assume the Gauss equation for $R^{[3]}$. Continuing in this way, we can formulate conditions

$$(II_k) \quad 1 \leq k \leq \ell - 1$$

$$(III_k) \quad 1 \leq k \leq \ell - 1$$

Finally, we can formulate

$$(IV) \quad R^{[\ell]} = 0.$$

Standard arguments about integrability conditions show that if the conditions (I), $\{(II_k)\}$, $\{(III_k)\}$, and (IV) hold, then the tensors $\mathcal{F}_0, \dots, \mathcal{F}_{\ell-1}$ on M come from an immersion of M into N .

PROBLEMS

1. Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a map with $\langle f(v), f(w) \rangle = \langle v, w \rangle$, where $\langle \cdot, \cdot \rangle$ is a non-degenerate inner product on \mathbb{R}^n . Show that

$$\langle f(\sum_i a_i e_i), f(e_j) \rangle = \langle \sum_i a_i f(e_i), f(e_j) \rangle$$

for all j , and conclude that f is linear.

2. Consider \mathbb{R}^{n+1} with the metric

$$-dx^0 \otimes dx^0 + dx^1 \otimes dx^1 + \cdots + dx^n \otimes dx^n.$$

(a) For the Levi-Civita connection (compare pg. II.342), the geodesics are the ordinary straight lines.

(b) If $g: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ is an isometry (with respect to this metric) with $g(0) = 0$ and $g_{*0} = \text{identity}$, then $g = \text{identity}$. [This can also be derived, as in Problem I-5, from an appropriate generalization of Corollary II.7-13.]

(c) If $f: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ is an isometry with $f(0) = 0$, then $f = f_{*0}$.

(d) Every isometry of \mathbb{R}^{n+1} is of the form $p \mapsto A(p) + q$ for $A \in O^1(n+1)$, and $q \in \mathbb{R}^{n+1}$.

3. Determine the geodesics of H^n by the same method used for S^n in Chapter I.9 (reflection through a 2-dimensional plane $P \subset \mathbb{R}^{n+1}$ is an isometry).

4. A **linear fractional transformation** is a map

$$z \mapsto \frac{az + b}{cz + d} \quad a, b, c, d \in \mathbb{C}, \quad ad - bc \neq 0,$$

of the extended complex plane $\mathbb{C} \cup \{\infty\}$ to itself.

(a) The set of all linear fractional transformations is a group under composition.

(b) For distinct z_1, z_2, z_3 , the transformation

$$z \mapsto \frac{z - z_2}{z - z_3} \bigg/ \frac{z_1 - z_2}{z_1 - z_3}$$

takes z_1 to 1, and z_2 to 0, and z_3 to ∞ .

(c) There is a linear fractional transformation taking any three distinct points $z_1, z_2, z_3 \in \mathbb{C} \cup \{\infty\}$ to any other three distinct points w_1, w_2, w_3 .

(d) If a linear fractional transformation keeps 1, 0, and ∞ fixed, then it is the identity.

(e) There is a unique linear fractional transformation taking z_1, z_2, z_3 to 1, 0, ∞ .

- (f) The transformation of part (c) is unique.
 (g) The linear fractional transformations which take the real axis to itself are precisely those with $a, b, c, d \in \mathbb{R}$.
 (h) The linear fractional transformations which take the upper half-plane onto itself are

$$f(z) = \frac{az + b}{cz + d},$$

$a, b, c, d \in \mathbb{R}$ and $ad - bc > 0$. We can then clearly assume that $ad - bc = 1$.

5. For distinct z_1, z_2, z_3 , the **cross ratio** (z, z_1, z_2, z_3) is defined as

$$(z, z_1, z_2, z_3) = \frac{z - z_2}{z - z_3} \bigg/ \frac{z_1 - z_2}{z_1 - z_3};$$

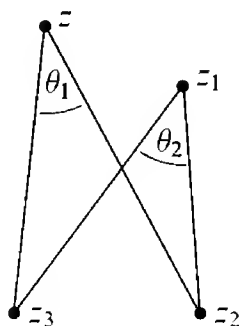
thus (z, z_1, z_2, z_3) is $f(z)$ where f is the linear fractional transformation taking z_1, z_2, z_3 to $1, 0, \infty$.

- (a) If g is a linear fractional transformation, then

$$(g(z), g(z_1), g(z_2), g(z_3)) = (z, z_1, z_2, z_3).$$

- (b) If $\theta = \arg w$ denotes an angle between the positive x -axis and the ray from 0 to w , so that $w = |w|e^{i\theta}$, then

$$\begin{aligned} \arg(z, z_1, z_2, z_3) &= \arg \frac{z - z_2}{z - z_3} - \arg \frac{z_1 - z_2}{z_1 - z_3} \\ &= \theta_1 - \theta_2 \quad \text{in the picture below.} \end{aligned}$$



Conclude that (z, z_1, z_2, z_3) is real if and only if z, z_1, z_2, z_3 lie on a circle or straight line.

- (c) A linear fractional transformation takes circles and straight lines into circles and straight lines.

6. In this problem we will use the notation on pages 319ff.

(a) The metric on the upper half-plane can be written

$$\langle \cdot, \cdot \rangle = \frac{dz \otimes d\bar{z}}{(\operatorname{Im} z)^2}.$$

(b) For the linear fractional transformation f of Problem 4(h), we have

$$\operatorname{Im} f(z) = \frac{\operatorname{Im} z}{|cz + d|^2}$$

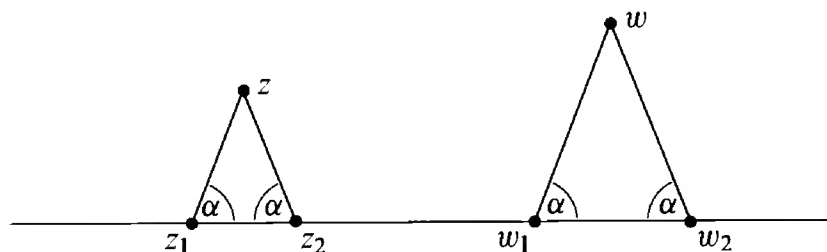
$$f_z = \frac{\partial}{\partial z} \left(\frac{az + b}{cz + d} \right) = \frac{1}{(cz + d)^2}$$

$$f^*(dz) = df = f_z dz + f_{\bar{z}} d\bar{z} = \frac{dz}{(cz + d)^2}$$

$$f^*(d\bar{z}) = d\bar{f} = \frac{d\bar{z}}{(c\bar{z} + d)^2}.$$

(c) Conclude that f is an isometry of the upper half-plane.

(d) There is such an isometry taking any given point z to any other. *Hint:* Consider the linear fractional transformation taking z_1, z_2, z in the figure below to w_1, w_2, w .



(e) In the B^2 model, the linear fractional transformations keeping $S = \text{boundary } B^2$ fixed are isometries, and there are such isometries taking any point to any other. Conclude that these isometries are all the orientation preserving isometries of B^2 , by noting that rotations about the origin are linear fractional transformations.

(f) The geodesic circles around 0 in B^2 are clearly ordinary circles. Conclude that all geodesic circles are ordinary circles, and that the same result holds in the upper half-plane. (The converse can be proved exactly as in the higher dimensional case.)

7. (a) In the upper half-plane, the distance between $z_1 = x + iy_1$ and $z_2 = x + iy_2$ is

$$d(z_1, z_2) = \left| \int_{y_1}^{y_2} \frac{dy}{y} \right| = \left| \log \frac{y_2}{y_1} \right| = |\log(z_0, z_1, z_2, z_3)|$$

where $z_0 = x$ and $z_3 = \infty$.

(b) Let z_1, z_2 be any two points of the upper half-plane and let the semi-circle through z_1, z_2 perpendicular to the x -axis meet the x -axis at z_0 and z_3 . Then

$$d(z_1, z_2) = |\log(z_0, z_1, z_2, z_3)|.$$

(c) Similarly, find the formula for $d(z_1, z_2)$ in B^2 .

8. (a) The only geodesic maps $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined on all of \mathbb{R}^n are the affine maps. *Hint:* Assume $f(0) = 0$, and recall the parallelogram law for addition, as on pg. III.211.

(b) Every geodesic map from S^{n+} to S^{n+} is of the form $\phi^{-1} \circ A \circ \phi$, where $\phi: S^{n+} \rightarrow \mathbb{R}^n$ is the standard geodesic map, and $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is affine.

9. In this Problem we will determine all geodesic maps $f: U \rightarrow V$ where U and V are open subsets of \mathbb{R}^n . We will use material from projective geometry—the reader is referred to Hartshorne [1] for all terms and theorems.* We need the fact that every $A = (a_{ij}) \in \text{GL}(n+1, \mathbb{R})$ determines a geodesic map $\bar{A}: \mathbb{P}^n \rightarrow \mathbb{P}^n$, and that every such map comes from some $A \in \text{GL}(n+1, \mathbb{R})$, unique up to multiplication by a real number. If we regard $\mathbb{R}^n \subset \mathbb{P}^n$, then the action of \bar{A} on \mathbb{R}^n is easily seen to be $\bar{A}(x^1, \dots, x^n) = (y^1, \dots, y^n)$, where

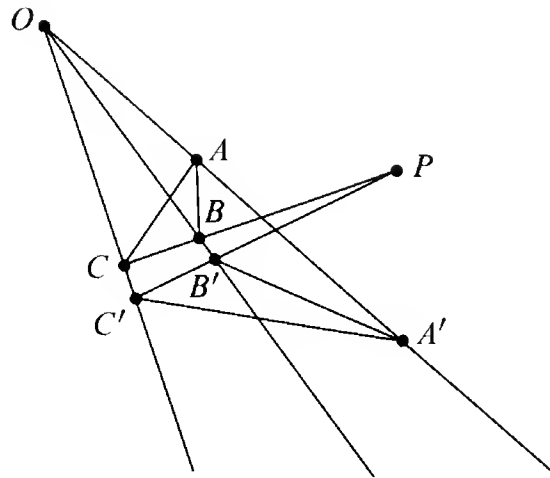
$$y^i = \frac{\sum_{j=1}^n a_{ij} x^j + a_{i,n+1}}{\sum_{j=1}^n a_{n+1,j} x^j + a_{n+1,n+1}}$$

(points where the denominator vanish go into the line at infinity). We will also use Desargue's Theorem and its converse (= its dual).

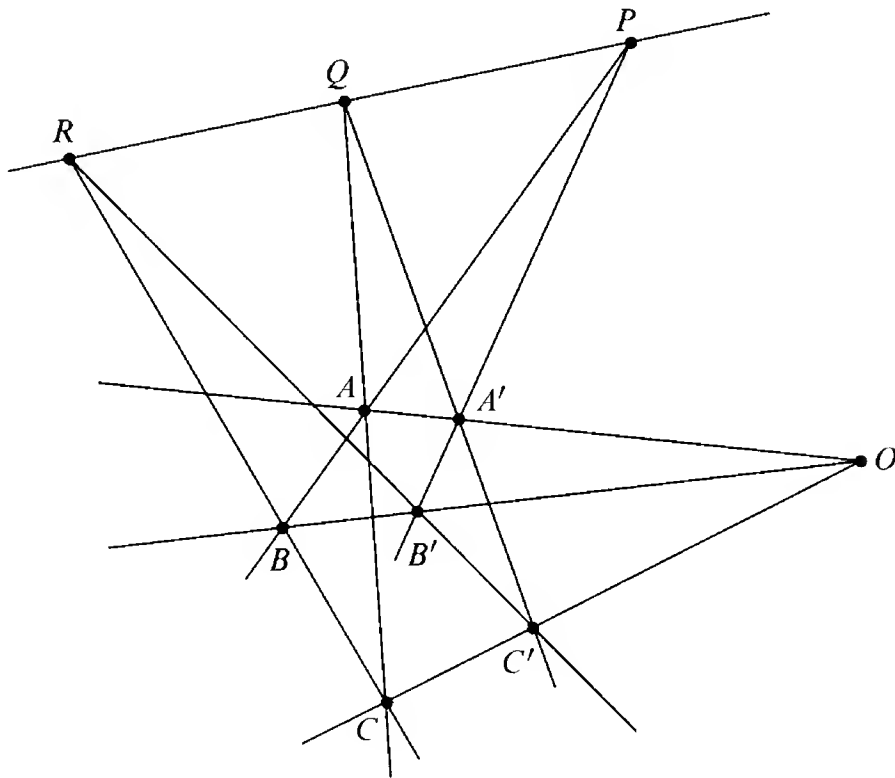
(a) Given any point $O \in \mathbb{R}^n$, and three lines l_1, l_2, l_3 through O which intersect U , show that there is a Desargue configuration with all other points in U . [*Hint:* In the figure at the top of the next page, points A, A' , and P are fixed, while B and B' , and C and C' , are chosen close together.] Conclude that the lines containing the $f(l_i \cap U)$ are concurrent. Thus show that there is a well-defined extension $\tilde{f}: \mathbb{P}^n \rightarrow \mathbb{P}^n$ with the property that if $P \in l_1 \cap l_2$ where l_1 and l_2 intersect U , then $\tilde{f}(P)$ is the intersection of the lines containing $f(l_1 \cap U)$ and $f(l_2 \cap U)$. Show also that \tilde{f} is one-one and onto.

(b) Let P, Q, R be three collinear points of \mathbb{P}^n . In the figure on the bottom of the next page, we first choose $A, A' \in U$, and then $B, B' \in U$, so that the lines

* For an analytic derivation see Scheffers [1; V.2. 429–432].



AA' and BB' intersect at a point $O \in U$. Show that we can also arrange for QA and RB to intersect at a point $C \in U$ and for RB' and QA' to intersect at a point $C' \in U$. Then show that AA' and BB' and CC' intersect at $O \in U$, so that we have a Desargue configuration with all points except P, Q, R in U . Conclude that $\tilde{f}(P)$, $\tilde{f}(Q)$, and $\tilde{f}(R)$ are collinear.



(c) Every geodesic map $f: U \rightarrow V$, where $U, V \in \mathbb{R}^n$ are open connected sets, is the restriction of some map \bar{A} for $A \in \text{GL}(n+1, \mathbb{R})$.

(d) Every geodesic map from H^n into H^n is of the form $\phi^{-1} \circ \bar{A} \circ \phi$, where $\phi: H^n \rightarrow B^n(1)$ is the standard geodesic map, and $\bar{A}: B^n(1) \rightarrow B^n(1)$ is a geodesic map which takes $B^n(1)$ into $B^n(1)$.

10. (a) Let $f: \mathbb{P}^2 \rightarrow \mathbb{P}^2$ be a geodesic map which takes a circle $\Sigma \subset \mathbb{R}^2 \subset \mathbb{P}^2$ into itself. Show that f is determined by knowing $f(P), f(Q), f(R)$ for distinct points $P, Q, R \in \Sigma$. *Hint:* Consider the tangent lines at P and Q , which intersect at some point S .

(b) Show that there is such an f for any given values of $f(P), f(Q), f(R)$. (You will need to use the fact that a conic is determined by 3 points and 2 tangents—see a book on projective geometry which treats conics.)

(c) Parts (a) and (b) show that the group of all geodesic maps $f: \mathbb{P}^2 \rightarrow \mathbb{P}^2$ with $f(\Sigma) = \Sigma$ has dimension 3. Using Problem 9, conclude that every geodesic map of H^2 onto itself is an isometry.

(d) Generalize to higher dimensions. Also consider the geodesic maps of S^n onto itself.

(e) Use the geodesic maps $H^n \rightarrow B^n(1)$ and $B^n(2) \rightarrow B^n(1)$ to describe an isometry between H^n and $B^n(2)$.

11. For vectors v_1, \dots, v_{m-1} in \mathbb{R}^m , we define $v_1 \times \dots \times v_{m-1}$ to be the unique vector with

$$\langle v_1 \times \dots \times v_{m-1}, w \rangle = \det \begin{pmatrix} v_1 \\ \vdots \\ v_{m-1} \\ w \end{pmatrix}$$

for all $w \in \mathbb{R}^m$.

(a) If $T: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is an orientation preserving isometry, then

$$T(v_1 \times \dots \times v_{m-1}) = T(v_1) \times \dots \times T(v_{m-1}).$$

(b) Show how to define $v_1 \times \dots \times v_{m-1}$ for v_1, \dots, v_{m-1} in an oriented m -dimensional vector space V with an inner product $\langle \cdot, \cdot \rangle$.

12. (a) Let c be an arclength parameterized curve in $(N, \langle \cdot, \cdot \rangle)$, with $\kappa_1, \dots, \kappa_{m-1} = 0$, and Frenet frame $\mathbf{v}_1, \dots, \mathbf{v}_{m-1}$. Using $\nu_r = \mathbf{v}_r$ as a trivialization of the normal bundle of image c , show that

$$\begin{aligned} \Pi^r(\mathbf{v}_1, \mathbf{v}_1) &= \kappa_1 \delta_2^r \\ \beta_r^s(\mathbf{v}_1) &= -\kappa_{r-1} \delta_{r-1}^s + \kappa_r \delta_{r+1}^s. \end{aligned}$$

Hence Π^r and β_r^s are expressible in terms of $\kappa_1, \dots, \kappa_{m-1}$, and conversely, $\kappa_1, \dots, \kappa_{m-1}$ are expressible in terms of the Π^r and β_r^s .

(b) Derive Corollary 4 from Theorem 20.

(c) Prove the assertion on page 51 by showing that $\phi \circ c = c$ for every curve $c: [0, 1] \rightarrow M$ with $c(0) = p$.

13. Let $M^n, \bar{M}^n \subset S^m \subset \mathbb{R}^{m+1}$, with corresponding covariant differentiations $\nabla, \nabla', \nabla''$ and $\bar{\nabla}, \bar{\nabla}', \bar{\nabla}''$ (as in the proof of Theorem 27). Let $\phi: M \rightarrow \bar{M}$ be an isometry, and $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ a bundle isomorphism covering ϕ between the normal bundles in S^m which preserves $\langle \cdot, \cdot \rangle$, s , and D . Let ν be the unit normal on S^m .

(a) The normal bundle $\text{Nor } M$ of M in \mathbb{R}^{m+1} has fibre $M_p^\perp = M_p^\perp \oplus \mathbb{R} \cdot \nu(p)$, and similarly for $\text{Nor } \bar{M}$. Define $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ extending $\tilde{\phi}$ by $\tilde{\phi}(\nu(p)) = \nu(\phi(p))$. Then $\tilde{\phi}$ is inner product preserving.

(b) The second fundamental form \mathbf{s} of M in \mathbb{R}^{m+1} is given by

$$\mathbf{s}(X, Y) = s(X, Y) + \langle X, Y \rangle \nu,$$

and similarly for \bar{M} .

(c) The normal connection \mathbf{D} of $\text{Nor } M$ is given by

$$\mathbf{D}_X \xi = D_X \xi \quad \xi \text{ a section of } \text{Nor } M$$

$$\mathbf{D}_X \nu = 0,$$

and similarly for \bar{M} .

(d) The bundle isomorphism $\tilde{\phi}$ preserves \mathbf{s} and \mathbf{D} , so there is a Euclidean motion $A: \mathbb{R}^{m+1} \rightarrow \mathbb{R}^{m+1}$ with $\phi = A|_M$ and $\tilde{\phi} = A_*|_{\text{Nor } M}$.

(e) From the action of $\tilde{\phi}$ on $\nu(p)$ conclude that A keeps 0 fixed, so that it also represents an isometry of S^m .

(f) Treat the case of two manifolds $M^n, \bar{M}^n \subset H^m$ similarly.

14. Let $(M, \langle \cdot, \cdot \rangle)$ be as in part (2) of Theorem 19, except with Gauss' Equation as on page 55, with $K_0 = 1$. Let $\varpi: \mathbf{E} \rightarrow M$ be the bundle whose fibre at p is $\varpi^{-1}(p) \oplus \mathbb{R}$, and extend $\{ \cdot, \cdot \}$ to a metric $\{ \cdot, \cdot \}$ by

$$\{(v, a), (w, b)\} = \{v, w\} + ab.$$

Define a symmetric section σ of $\text{Hom}(TM \times TM, \mathbf{E})$ by

$$\sigma(X, Y) = (\sigma(X, Y), \langle X, Y \rangle),$$

and define a connection δ on \mathbf{E} compatible with $\{ \cdot, \cdot \}$ by

$$\begin{aligned} \delta_X \xi &= \delta_X \xi & \xi \text{ a section of } E \\ \delta_X \zeta &= 0 & \text{where } \zeta \text{ is the section} \\ & & \zeta(p) = (0, 1) \in \varpi^{-1}(p) \oplus \mathbb{R}. \end{aligned}$$

- (a) Gauss' equation, in the form with $K_0 = 0$, holds for σ .
- (b) The Codazzi-Mainardi equations hold for $\tilde{\nabla}\sigma$.
- (c) The Ricci equations hold for $R_\delta, \sigma, \mathbf{A}_\xi$.
- (d) Let $f: M \rightarrow \mathbb{R}^{m+1}$ be the isometric immersion given by Theorem 19, for $M, \mathbf{E}, \{ \cdot, \cdot \}, \sigma, \delta$. Regard f as an imbedding (by working locally), and let v be the vector field $\tilde{f}(\zeta)$ on $f(M)$. Then for all tangent vectors X, Y of $f(M)$ we have

$$\langle s(X, Y), v \rangle = \langle X, Y \rangle \implies \nabla'_X v = -X.$$

- (e) Let $p \in f(M)$ be a fixed point. Changing f by a translation, we can assume that $v(p) = -p$ (identifying tangent vectors of \mathbb{R}^{m+1} with elements of \mathbb{R}^{m+1} , as usual). Let $c: [0, 1] \rightarrow f(M)$ be a curve with $c(0) = p$. Then

$$\frac{dv(c(t))}{dt} = -c'(t),$$

and consequently $v(c(t)) = -c(t)$ for all t . Conclude that $f(M) \subset S^m$.

- (f) Treat the case $K_0 = -1$ similarly.

15. The Lie algebra $\mathfrak{gl}(m, \mathbb{R})$ has as a basis the matrices E_α^β which have zeros everywhere except for a 1 in *column* α and *row* β , so that

$$(E_\alpha^\beta)^\rho_\sigma = \delta_\alpha^\rho \delta_\sigma^\beta.$$

Let $\{\psi_\alpha^\beta\}$ be the dual basis, and let $\tilde{\psi}_\alpha^\beta$ be the left invariant 1-forms on $\text{GL}(m, \mathbb{R})$ which extend the ψ_α^β .

- (a) Show that

$$d\tilde{\psi}_\alpha^\beta = - \sum_{\gamma=1}^m \tilde{\psi}_\gamma^\beta \wedge \tilde{\psi}_\alpha^\gamma,$$

either by computing the brackets of the E_α^β and using the first equation on pg. I.396, or, more easily, by using the last equation on pg. I.404.

(b) The Lie algebra $\mathfrak{o}(m)$ has as a basis the matrices $E_\alpha^\beta - E_\beta^\alpha$, $\alpha < \beta$. The dual basis is

$$(l) \quad \phi_\alpha^\beta = \frac{\psi_\alpha^\beta - \psi_\beta^\alpha}{2}, \quad \alpha < \beta.$$

Define $\phi_\alpha^\beta = -\phi_\beta^\alpha$ for $\alpha > \beta$ and $\phi_\alpha^\alpha = 0$. Note that equation (l) still holds. Verify that we now have

$$d\tilde{\phi}_\alpha^\beta = -\sum_{\gamma=1}^m \tilde{\phi}_\gamma^\beta \wedge \tilde{\phi}_\alpha^\gamma.$$

(c) Derive Theorem 19 as a consequence of Theorems I.10-17 and I.10-18.

16. Use Problem I.7-14(a) to show that the even powers of λ in the characteristic polynomial $\chi(\lambda)$ of A can be expressed in terms of the determinants of the 2×2 submatrices of A .

17. For a hypersurface $M \subset \mathbb{R}^{n+1}$, generalize Proposition 2-6 so as to express the $(n+1)^{\text{st}}$ fundamental form in terms of the first n fundamental forms and the elementary symmetric curvatures.

18. For an immersion $f: M^n \rightarrow \mathbb{R}^{n+1}$ with normal map $N_f = \nu \circ f$, show that we still have

$$\text{III}_f = \text{I}_{N_f} = -\text{II}_{N_f}.$$

19. Let c be a curve in a hypersurface $M \subset N$ of a manifold of constant curvature K_0 , and let X be a vector field of N along M . Then $\nabla'_{c'(s)} X$ is always a multiple of $c'(s)$ if and only if the ruled surface $\{\exp_{c(s)} tX(c(s))\}$ has constant intrinsic curvature K_0 .

20. Let $\sigma: S^n - \{\text{north pole}\} \rightarrow \mathbb{R}^n$ be the version of stereographic projection in which S^n denotes the standard unit sphere around 0.

(a) For this σ we have

$$\sigma(p) = \left(\frac{p^1}{1 - p^{n+1}}, \dots, \frac{p^n}{1 - p^{n+1}} \right)$$

$$\sigma^{-1}(y) = \left(\frac{2y^1}{1 + \sum_i (y^i)^2}, \dots, \frac{2y^n}{1 + \sum_i (y^i)^2}, \frac{\sum_i (y^i)^2 - 1}{1 + \sum_i (y^i)^2} \right).$$

(b) Let $c: [0, 2\pi] \rightarrow \mathbb{R}^n$ be a curve, parameterized proportionally to arclength, which goes once around a circle centered at 0 and passing through y , so that c' has squared length $|y|^2$. Then $(\sigma^{-1} \circ c)'$ has squared length

$$\frac{4|y|^2}{[1 + |y|^2]^2}.$$

Thus σ^{-1}_* multiplies lengths of tangent vectors at y by $2/(1 + |y|^2)$.

21. Let $\sigma: S^2 \rightarrow \mathbb{C} \cup \{\infty\}$ be stereographic projection, where S^2 is the standard unit sphere around $(0, 0, 0)$.

(a) If R_θ is rotation of S^2 through an angle of θ around the z -axis, then $\sigma \circ R_\theta \circ \sigma^{-1}: \mathbb{C} \cup \{\infty\} \rightarrow \mathbb{C} \cup \{\infty\}$ is just $z \mapsto e^{i\theta}z$.

(b) If R'_θ is rotation through an angle of θ around the y -axis, calculate that $\sigma \circ R'_\theta \circ \sigma^{-1}$ is

$$z \mapsto \frac{(1 + \cos \theta)z - \sin \theta}{(\sin \theta)z + (1 + \cos \theta)}.$$

(c) The group $SO(3)$ is generated by all R_θ and R'_θ . (A direct proof can be given, or one can note that $SO(3)$ is 3-dimensional, and the R_θ and R'_θ do not commute.) The group of all 4×4 complex matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ satisfying the conditions on page 109 is also 3-dimensional. Conclude that this group is precisely the group of all $\sigma \circ A \circ \sigma^{-1}$ for $A \in O(3)$.

22. Consider $B^2(2)$, with the metric on page 7. From pg. II.301 we see that the geodesic circle of radius r is given by

$$c(\theta) = 2 \tanh \frac{r}{2} (\cos \theta, \sin \theta) \quad 0 \leq \theta \leq 2\pi.$$

(a) Calculate that

$$|c'(\theta)| = \sinh r.$$

(b) Then verify the formula for I given on page 118.

23. (a) For a coordinate system u, v on a 2-dimensional Riemannian manifold, show that the formula on page 132 can be written

$$\Delta f = \frac{1}{W} \left\{ \frac{\partial}{\partial u} \left(\frac{G \frac{\partial f}{\partial u} - F \frac{\partial f}{\partial v}}{W} \right) + \frac{\partial}{\partial v} \left(\frac{E \frac{\partial f}{\partial v} - F \frac{\partial f}{\partial u}}{W} \right) \right\},$$

where $W = \sqrt{EG - F^2}$.

(b) If (u, v) is isothermal (this means that $E = G$ and $F = 0$; compare pg. II.297), then

$$\Delta f = \frac{1}{E} \left(\frac{\partial^2 f}{\partial u^2} + \frac{\partial^2 f}{\partial v^2} \right).$$

(c) A coordinate system (x, y) on a 2-dimensional Riemannian manifold is isothermal if and only if $\Delta_1 x = \Delta_1 y$ and $\Delta_1(x, y) = 0$.

(d) If (x, y) is an isothermal coordinate system, then $\Delta x = \Delta y = 0$.

(e) If $\Delta x = 0$, then there is locally a function y with

$$dy = \frac{F \frac{\partial x}{\partial u} - E \frac{\partial x}{\partial v}}{W} du + \frac{G \frac{\partial x}{\partial u} - F \frac{\partial x}{\partial v}}{W} dv$$

(here E, F, G are the components of $\langle \cdot, \cdot \rangle$ in the (u, v) coordinate system). The functions x and y satisfy $\Delta_1 x = \Delta_1 y$ and $\Delta_1(x, y) = 0$, so (x, y) is an isothermal coordinate system.

24. Let h be a function on a 2-dimensional Riemannian manifold such that the sets $h = \text{constant}$ give a foliation of the manifold.

(a) Suppose that there is an isothermal coordinate system (x, y) such that one family of parameter curves lie along the curves $h = \text{constant}$; thus $x = f \circ h$ for some function f . Use Problem 23 to show that

$$\Delta h \cdot (f' \circ h) + \Delta_1 h \cdot (f'' \circ h) = 0.$$

Hence $\Delta h / \Delta_1 h$ is some function composed with h .

(b) Conversely, if $\Delta h / \Delta_1 h = F \circ h$ for some function F , and we set $x = f \circ h$ for

$$f' = e^{-\int F},$$

then $\Delta x = 0$, and the function y of Problem 23(e) satisfies

$$\Delta_1 y = e^{-2 \int F} \Delta_1 h.$$

(c) So

$$\langle \cdot, \cdot \rangle = \frac{1}{\Delta_1 h} (dh \otimes dh + e^{2 \int F} dy \otimes dy).$$

(d) If we have equations (a) and (b) on page 155, then the corresponding metrics are

$$\begin{aligned} & \frac{1}{f \circ K} (dK \otimes dK + e^{2 \int g/f} dy \otimes dy) \\ & \frac{1}{f \circ \bar{K}} (d\bar{K} \otimes d\bar{K} + e^{2 \int g/f} d\bar{y} \otimes d\bar{y}). \end{aligned}$$

So there is a one-parameter family of isometries between the surfaces.

(e) There is a function x with

$$dx \otimes dx = \frac{1}{f \circ K} dK \otimes dK,$$

and similarly for \bar{x} . Hence, each surface is isometric to a surface of revolution (see formula (4) on pg. III.158).

25. Let V and W be two inner product spaces of the same dimension. Let $\{v_\rho\}$ be an indexed set of (not necessarily distinct) vectors which span V , and let $\{\bar{v}_\rho\}$ span W . Suppose that

$$\langle v_\rho, v_\sigma \rangle = \langle \bar{v}_\rho, \bar{v}_\sigma \rangle \quad \text{for all } \rho, \sigma.$$

Show that $\sum_\rho c_\rho v_\rho = 0 \implies \sum_\rho c_\rho \bar{v}_\rho = 0$, and conclude that there is a unique inner product preserving isomorphism $V \rightarrow W$ which takes v_ρ to \bar{v}_ρ .

CHAPTER 8

THE SECOND VARIATION

In this chapter we return to the study of the calculus of variations, and introduce an important (essentially classical) construction, which has surprisingly significant consequences for differential geometry. Recall that the calculus of variations was first invoked in order to find paths which locally minimize the length function L for a Riemannian manifold M . In the course of our investigations we found that the energy function was more convenient to work with, and that the critical paths for the length function are precisely the same as those for the energy function, except that the latter are necessarily parameterized proportionally to arclength. These critical points for E are, of course, the geodesics on M , and at present we know only that *sufficiently small* pieces of geodesics are paths of minimal length.

We now want to develop conditions which determine when a given geodesic is, in its entirety, a path of smaller length than nearby paths. We recall one fact from Problem I.9-31: For a piecewise C^∞ curve $\gamma: [a, b] \rightarrow M$ we always have

$$[L_a^b(\gamma)]^2 \leq (b - a)E_a^b(\gamma),$$

with equality precisely when γ is parameterized proportionally to arclength. From this it is easy to see that a geodesic γ has minimal *length* among all nearby paths between $\gamma(a)$ and $\gamma(b)$ precisely when it has minimal *energy* among all such paths. Thus we lose no information by restricting all our considerations to the energy function E .

We begin with a brief summary of the results which we already have. Consider a piecewise C^∞ path $\gamma: [a, b] \rightarrow M$ and a piecewise C^∞ variation $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ of γ . We define

$$\begin{aligned} W(t) &= \frac{\partial \alpha}{\partial u}(0, t) && \text{the "variation vector field"} \\ V(t) &= \frac{d\gamma}{dt} && \text{the "velocity vector field of } \gamma\text{"} \\ A(t) &= \frac{D}{dt} V(t) && \text{the "acceleration vector field of } \gamma\text{",} \end{aligned}$$

and if $a = t_0 < \cdots < t_N = b$ includes all discontinuity points of V , we set

$$\begin{aligned}\Delta_{t_i} V &= V(t_i^+) - V(t_i^-) \quad i = 1, \dots, N-1 \\ \Delta_{t_0} V &= V(t_0^+) \\ \Delta_{t_N} V &= -V(t_N^-).\end{aligned}$$

We then have the following formula (Theorem II.6-14) for the “first variation” of E :

$$\left. \frac{dE(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_a^b \langle W(t), A(t) \rangle dt - \sum_{i=0}^N \langle W(t_i), \Delta_{t_i} V \rangle;$$

for variations keeping the endpoints fixed, the sum can be written from 1 to $N-1$. From this formula we found that γ is a geodesic ($A(t) = 0$) if and only if γ is a critical point for E .

Recall that if $f: M \rightarrow \mathbb{R}$ is a real-valued function, then $f_*: M_p \rightarrow \mathbb{R}_{f(p)}$ may be determined as follows. Given $X_p \in M_p$, we choose a path $c: (-\varepsilon, \varepsilon) \rightarrow M$ with $c'(0) = X_p$; then

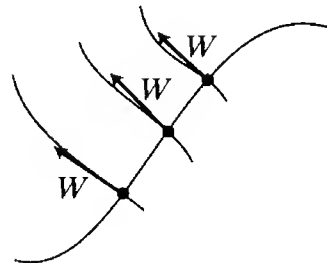
$$f_*(X_p) = \text{tangent vector of } f \circ c \text{ at } 0 = \left. \frac{df(c(u))}{du} \right|_{u=0} \cdot \left. \frac{d}{dt} \right|_{f(p)}.$$

This suggests some notation which is exactly analogous, except that we will be sloppy and throw away the uninteresting d/dt term. For any piecewise C^∞ vector field W along γ , we define

$$E_*(W) = \left. \frac{dE(\bar{\alpha}(u))}{du} \right|_{u=0},$$

where α is some piecewise C^∞ variation of γ with W as its variation vector field. The first variation formula shows that the right side depends only on W , so that $E_*(W)$ is really well-defined; the formula also shows that E_* is linear. Perhaps we should explicitly make the observation that any piecewise C^∞ vector field W is the variation vector field of some α ; for example, we can take

$$\alpha(u, t) = \exp u \cdot W(t).$$



As this example shows, we can even arrange for α to be a variation keeping endpoints fixed if $W(a) = W(b) = 0$. The notation $E_*(W)$ suggests that piecewise C^∞ vector fields W along γ may be thought of as “tangent vectors” to the curve γ . Actually, it will be convenient to restrict this terminology to those W which vanish at a and b . So if Ω denotes the set of all piecewise C^∞ paths $\gamma: [a, b] \rightarrow M$ between two fixed points p and q , we will define Ω_γ , the “tangent space of Ω at γ ”, to be the vector space

$$\Omega_\gamma = \{W : W \text{ is a piecewise } C^\infty \text{ vector field along } \gamma \text{ with } W(a) = W(b) = 0\}.$$

We know that if $E: \Omega \rightarrow \mathbb{R}$ has a minimum, or even a local minimum, at γ , then γ must be a geodesic, so $E_*: \Omega_\gamma \rightarrow \mathbb{R}$ must be 0. This is a *necessary* condition, analogous to the *necessary* condition $D_i f(x) = 0$ for a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ to have a local maximum or minimum at $x \in \mathbb{R}^n$. We also want to find *sufficient* conditions for a geodesic γ to be a minimum for E ; as a guide, we will first recall what is known in the case of functions $f: \mathbb{R}^n \rightarrow \mathbb{R}$.

In the one variable case, there are very easy sufficient conditions for a function $f: \mathbb{R} \rightarrow \mathbb{R}$ to have a local maximum or minimum:

- (1) If $f'(x) = 0$ and $f''(x) > 0$, then f has a (strict) local minimum at x .
- (2) If $f'(x) = 0$ and $f''(x) < 0$, then f has a (strict) local maximum at x .

To prove (1), for example, we simply note that if $f'(x) = 0$ and $f''(x) > 0$, then we must have $f'(x+h) > 0$ for small $h > 0$, and $f'(x+h) < 0$ for small $h < 0$. So f is strictly decreasing in some interval $(x-\varepsilon, x]$, and strictly increasing on some interval $[x, x+\varepsilon)$. We also obtain, automatically, the following partial converses:

- (1') If f has a local minimum at x , and $f''(x)$ exists, then $f''(x) \geq 0$.
- (2') If f has a local maximum at x , and $f''(x)$ exists, then $f''(x) \leq 0$.

[Proof of (1'): If we had $f''(x) < 0$, then f would have a strict local maximum at x , by (2), contradicting the hypothesis that it has a local minimum at x .]

For functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}$, the situation becomes more complicated. We certainly cannot expect to conclude that a critical point x of f is a local minimum simply because

$$D_{1,1}f(x) > 0 \quad \text{and} \quad D_{2,2}f(x) > 0;$$

this condition merely implies that x is a local minimum for f along the lines through x which are parallel to one of the axes. We would need the much

stronger condition (Problem 1) that every second order directional derivative of f is positive,

$$\left. \frac{d^2}{dt^2} \right|_{t=0} f(x + tv) > 0 \quad \text{all } v \in \mathbb{R}^2.$$

If we use *mixed* partial derivatives, then we have a simple sufficient condition that a critical point x be *either* a (strict) local maximum *or* a (strict) local minimum, namely

$$(I) \quad \det \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right) > 0.$$

We have essentially already proved this in Chapter 2, for this inequality is exactly the condition that x be an elliptic point of the surface $\{(x_1, x_2, f(x_1, x_2))\}$, and therefore lie on one side of its tangent plane at x ; this tangent plane is just the (x_1, x_2) -plane, since x is a critical point. If condition (I) is satisfied, we can then distinguish between a local maximum and a local minimum merely by examining the sign of $\partial^2 f / \partial (x_1)^2$ at x . If, instead of condition (I), we have

$$(II) \quad \det \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right) < 0,$$

then x definitely is *not* either a local maximum or minimum for f . This also follows from the considerations of Chapter 2, for in this case the surface $\{(x_1, x_2, f(x_1, x_2))\}$ lies on both sides of its tangent plane. When the determinant is 0, we are in the borderline case where no conclusions can be drawn. Essentially the same considerations hold for functions $f: \mathbb{R}^n \rightarrow \mathbb{R}$, except that it is no longer so easy to find out if the eigenvalues of

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)$$

all have the same sign, which is precisely the condition that f have either a local maximum or a local minimum at x .

Notice that the analogues of (1') and (2') require no modification: If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ has a local minimum at x , then surely

$$\left. \frac{d^2}{dt^2} \right|_{t=0} f(x + tv) \geq 0$$

for all $v \in \mathbb{R}^n$ for which this limit exists. In fact, if the opposite equality held for some $v \in \mathbb{R}^n$, then f would have a strict local maximum at x along the line $\{x + tv: t \in \mathbb{R}\}$.

Our aim now is to see what information we can get when we generalize these considerations of elementary calculus, and examine the second derivative $d^2E(\bar{\alpha}(u))/du^2(0)$, for all variations α of a geodesic $\gamma: [a, b] \rightarrow M$; classically, this second derivative was called the “second variation” of E . Our remarks about n -dimensional calculus might suggest that it would be even more useful to consider “mixed partial derivatives”, and even if they don’t suggest it, mixed partial derivatives do turn out to be the thing to look at. We first define a **2-parameter variation** α of γ to be a function

$$\alpha: U \times [a, b] \rightarrow M,$$

for some neighborhood U of $0 \in \mathbb{R}^2$, such that

- (1) $\alpha(0, t) = \gamma(t)$
- (2) there is a partition $a = t_0 < \cdots < t_N = b$ of $[a, b]$ so that α is C^∞ on each $U \times [t_{i-1}, t_i]$.

We say that α is a variation **keeping endpoints fixed** if

- (3) For all $u \in U$, we have

$$\begin{aligned}\alpha(u, a) &= \gamma(a) \\ \alpha(u, b) &= \gamma(b).\end{aligned}$$

As before, we let $\bar{\alpha}(u)$ be the path $t \mapsto \alpha(u, t)$. A 2-parameter variation α of γ gives rise to two “variation vector fields” W_1 and W_2 along γ , defined by

$$W_i(t) = \frac{\partial \alpha}{\partial u_i}(0, 0, t).$$

Notice that the W_i may be only piecewise C^∞ vector fields along γ even if γ itself is everywhere C^∞ .

1. THEOREM (SECOND VARIATION FORMULA). Let $\gamma: [a, b] \rightarrow M$ be a geodesic, with velocity vector field $V(t) = d\gamma/dt$, and let $\alpha: U \times [a, b] \rightarrow M$ be a 2-parameter variation of γ , with variation vector fields

$$W_i(t) = \frac{\partial \alpha}{\partial u_i}(0, 0, t).$$

Choose $a = t_0 < \cdots < t_N = b$ to include all discontinuity points of DW_1/dt , and let

$$\begin{aligned}\Delta_{t_i} \frac{DW_1}{dt} &= \frac{DW_1}{dt}(t_i^+) - \frac{DW_1}{dt}(t_i^-) \quad i = 1, \dots, N-1 \\ \Delta_{t_0} \frac{DW_1}{dt} &= \frac{DW_1}{dt}(t_0^+) \\ \Delta_{t_N} \frac{DW_1}{dt} &= -\frac{DW_1}{dt}(t_N^-).\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_1 \partial u_2} \Big|_{(u_1, u_2) = (0, 0)} &= - \int_a^b \left\langle W_2(t), \frac{D^2 W_1}{dt^2} + R(W_1(t), V(t))V(t) \right\rangle dt \\ &\quad - \sum_{i=0}^N \left\langle W_2(t_i), \Delta_{t_i} \frac{DW_1}{dt} \right\rangle.\end{aligned}$$

(When α is a variation keeping endpoints fixed, the sum can be written from 1 to $N-1$.)

PROOF. By the first variation formula (Theorem II.6-14), we have

$$\frac{\partial E(\bar{\alpha}(u))}{\partial u_2} \Big|_{u_2=0} = - \int_a^b \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{dt} \frac{\partial \alpha}{\partial t} \right\rangle dt - \sum_{i=0}^N \left\langle \frac{\partial \alpha}{\partial u_2}, \Delta_{t_i} \frac{\partial \alpha}{\partial t} \right\rangle,$$

where all terms on the right side are to be evaluated at $(t, u_1, 0)$. So

$$\begin{aligned}(1) \quad \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_1 \partial u_2} \Big|_{u_2=0} &= - \int_a^b \left\langle \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt - \int_a^b \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1} \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt \\ &\quad - \sum_{i=0}^N \left\langle \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial u_2}, \Delta_{t_i} \frac{\partial \alpha}{\partial t} \right\rangle - \sum_{i=0}^N \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1} \Delta_{t_i} \frac{\partial \alpha}{\partial t} \right\rangle.\end{aligned}$$

Now when $u_1 = 0$ we have

$$\frac{D}{\partial t} \frac{\partial \alpha}{\partial t}(t, 0, 0) = 0 \quad \text{and} \quad \Delta_{t_i} \frac{\partial \alpha}{\partial t}(t, 0, 0) = 0,$$

since $t \mapsto \alpha(t, 0, 0) = \gamma(t)$ is a geodesic. So the first and third terms on the right side of equation (1) are zero for $u_1 = 0$. After a simple manipulation with

the fourth term we then have

$$(2) \quad \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_1 \partial u_2} \Big|_{(u_1, u_2)=(0,0)} = - \int_a^b \left\langle W_2(t), \frac{D}{\partial u_1} \frac{D}{\partial t} V \right\rangle dt \\ - \sum_{i=0}^N \left\langle W_2(t_i), \Delta_{t_i} \frac{DW_1}{dt} \right\rangle,$$

where all terms on the right are now evaluated at $(t, 0, 0)$. Now we can use Proposition II.6-10 to write

$$\frac{D}{\partial u_1} \frac{D}{\partial t} V - \frac{D}{\partial t} \frac{D}{\partial u_1} V = R \left(\frac{\partial \alpha}{\partial u_1}, \frac{\partial \alpha}{\partial t} \right) V = R(W_1, V) V.$$

Moreover, Proposition II.6-9 gives us

$$\frac{D}{\partial u_1} V = \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial t} = \frac{D}{\partial t} \frac{\partial \alpha}{\partial u_1} = \frac{D}{dt} W_1,$$

so we have

$$\frac{D}{\partial u_1} \frac{D}{\partial t} V = \frac{D^2 W_1}{dt^2} + R(W_1, V) V.$$

Substituting into (2), we obtain the desired result. ♦

Suppose we are given two piecewise C^∞ vector fields W_1 and W_2 along a geodesic $\gamma: [a, b] \rightarrow M$. We can always find at least one variation α with these as variation vector fields, namely

$$\alpha(u_1, u_2, t) = \exp[u_1 W_1(t) + u_2 W_2(t)].$$

Extending the notation introduced previously, we define

$$E_{**}(W_1, W_2) = \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_1 \partial u_2} \Big|_{(u_1, u_2)=(0,0)},$$

for any variation α with variation vector fields W_1 and W_2 ; the second variation formula shows that $E_{**}(W_1, W_2)$ does not depend on the choice of α . The notation $E_{**}(W_1, W_2)$ is used *only* when W_1 and W_2 are vector fields along a *geodesic*; otherwise the second derivative will depend on the choice of α (compare pg. I.161 and Problem I.5-17). It is clear from the second variation formula that E_{**} is bilinear. It is also true that E_{**} is symmetric, $E_{**}(W_1, W_2) = E_{**}(W_2, W_1)$; this is not at all clear from the second variation formula, but it

follows immediately from the fact that $E(\bar{\alpha}(u))$ is a C^∞ function of u , and consequently

$$\frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_1 \partial u_2} = \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u_2 \partial u_1}.$$

The second variation formula reveals a hitherto unsuspected significance of curvature, and turns out to be responsible for many of the deeper consequences which we will be able to draw from assumptions about the curvature of M . We begin the program which will uncover these results by formulating questions about local minima for E in terms of E_{**} . Notice that if $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ is a 1-parameter variation of γ , and we define a 2-parameter variation β by

$$\beta(u_1, u_2, t) = \alpha(u_1 + u_2, t),$$

then

$$\left. \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u^2} \right|_{u=0} = \left. \frac{\partial^2 E(\bar{\beta}(u))}{\partial u_1 \partial u_2} \right|_{(u_1, u_2)=(0,0)}.$$

If γ has variation vector field W , then β clearly has variation vector fields $W_1 = W_2 = W$. Consequently,

$$\left. \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u^2} \right|_{u=0} = E_{**}(W, W).$$

Thus, if γ is going to be a local minimum for energy, then we must have $E_{**}(W, W) \geq 0$ for all $W \in \Omega_\gamma$. Briefly expressed:

If γ is a local minimum, then E_{**} is positive semi-definite.

We also *hope* that γ actually will be a local minimum whenever we have the strict inequality $E_{**}(W, W) > 0$ for all non-zero $W \in \Omega_\gamma$. Briefly expressed:

If E_{**} is positive definite, then we *hope* that γ is a local minimum.

Our approach to this problem will be somewhat roundabout; we first investigate the vector fields $W \in \Omega_\gamma$ which satisfy $E_*(W, W) = 0$ for all $W \in \Omega_\gamma$, and hence represent something of a borderline between positive definiteness and positive semi-definiteness.

A piecewise C^∞ vector field W along γ is called a **Jacobi field** if it satisfies the “Jacobi equation”

$$\frac{D^2 W}{dt^2} + R(W, V)V = 0, \quad V = d\gamma/dt.$$

In a coordinate system this equation becomes a *linear* second order differential equation. Or, if we choose parallel vector fields Y_1, \dots, Y_n along γ which are orthonormal at 0, and hence orthonormal everywhere along γ , and set $W(t) = \sum_i f^i(t) Y_i(t)$, then our equation becomes

$$0 = \frac{d^2 f^i}{dt^2} + \sum_{j=1}^n a_j^i(t) f^j(t) \quad i = 1, \dots, n,$$

where $a_j^i = \langle R(Y_j, V)V, Y_i \rangle$. The solutions of this equation are everywhere C^∞ and, since the equation is linear, every solution can be defined on all of γ . It is also clear from the linearity of the equation that the set of all Jacobi fields W along γ forms a vector space. The dimension of this vector space is $2n$, since each Jacobi field W is determined by its initial conditions

$$W(0), \frac{DW}{dt}(0) \in M_{\gamma(0)}.$$

2. PROPOSITION. Let $\gamma: [a, b] \rightarrow M$ be a geodesic and let $W \in \Omega_\gamma$. Then W is a Jacobi field if and only if

$$E_{**}(W, W_2) = 0$$

for all $W_2 \in \Omega_\gamma$.

PROOF. If $W \in \Omega_\gamma$ is a Jacobi field, then the second variation formula shows immediately that

$$E_{**}(W, W_2) = - \int_a^b \langle W_2, 0 \rangle dt - \sum_{i=1}^{N-1} \langle W_2(t_i), 0 \rangle = 0.$$

Conversely, suppose that $W \in \Omega_\gamma$ and that $E_{**}(W, W_2) = 0$ for all $W_2 \in \Omega_\gamma$. Choose $a = t_0 < \dots < t_N = b$ so that each $W|_{[t_{i-1}, t_i]}$ is smooth, and let $f: [a, b] \rightarrow [0, 1]$ be a C^∞ function with $f(t_i) = 0$ and $f > 0$ otherwise. If we define

$$W_2 = f \cdot \left(\frac{D^2 W}{dt^2} + R(W, V)V \right),$$

then

$$0 = E_{**}(W, W_2) = - \int_a^b f \cdot \left\| \frac{D^2 W}{dt^2} + R(W, V)V \right\| dt - \sum_{i=0}^N \left\langle 0, \Delta_{t_i} \frac{DW}{dt} \right\rangle.$$

This implies that

$$(1) \quad \frac{D^2 W}{dt^2} + R(W, V)V = 0 \quad \text{on each } (t_{i-1}, t_i),$$

so each $W|_{[t_{i-1}, t_i]}$ is a Jacobi field.

Next choose W_2 to be any vector field along γ with $W_2(a) = W_2(b) = 0$ and $W_2(t_i) = \Delta_{t_i} DW/dt$ for $i = 1, \dots, N-1$. Then by (1) we have

$$0 = E_{**}(W, W_2) = - \int_a^b \langle W, 0 \rangle dt - \sum_{i=1}^N \left\| \Delta_{t_i} \frac{DW}{dt} \right\|^2,$$

so each $\Delta_{t_i} DW/dt = 0$. This means that the Jacobi fields $W|_{[t_{i-1}, t_i]}$ for two consecutive intervals have the same values of DW/dt on the intersection of the intervals. Since a Jacobi field is determined by its initial values, this shows that W is actually a Jacobi field on all of γ . ♦

Notice that there may not exist any non-trivial Jacobi fields W along γ which vanish at both a and b (indeed we hope to find conditions under which E_{**} is positive definite, which certainly excludes the possibility of non-zero Jacobi fields). When there is a non-zero Jacobi field W along γ with $W(a) = W(b) = 0$, we say that a and b are **conjugate values** along γ , and we define the **multiplicity** of a and b as conjugate values to be the dimension of the vector space consisting of all such Jacobi fields. We also say that $\gamma(a)$ and $\gamma(b)$ are **conjugate points** of γ , but this terminology is ambiguous when γ has self-intersections.

Since a Jacobi field W is determined by its initial values $W(a), DW/dt(a)$ at any point a , the multiplicity of two conjugate values a and b is clearly $\leq n$. Actually, it is always $\leq n-1$. To prove this, we just have to produce a Jacobi field along γ which is 0 at a but nowhere else. The vector field $W(t) = (t-a)V(t)$ has this property; it is a Jacobi field because

$$\begin{aligned} \frac{DW}{dt} &= V(t) + (t-a)\frac{DV}{dt} = V(t), \\ \frac{D^2 W}{dt^2} + R(W, V)V &= \frac{DV}{dt} + (t-a)R(V, V)V = 0. \end{aligned}$$

More generally, we have

3. PROPOSITION. Let γ be a geodesic, with velocity vector field $V = d\gamma/dt$.

(1) The vector field fV along γ is a Jacobi field if and only if f is linear.

- (2) Every Jacobi field W along γ can be written uniquely as $fV + W^\perp$, where f is linear and W^\perp is a Jacobi field perpendicular to γ .
- (3) If a Jacobi field W along γ is perpendicular to γ at two points a and b , then W is perpendicular to γ everywhere. In particular, if $W(a) = W(b) = 0$, then W is perpendicular to γ everywhere.

PROOF. (1) If $W = fV$, then $D^2W/dt^2 = f''V$, so the Jacobi equation for W is

$$0 = \frac{D^2W}{dt^2} + R(W, V)V = f''V + fR(V, V)V = f''V.$$

(2) Given a Jacobi field W along γ , we can write $W = fV + W^\perp$ for some f and some vector field W^\perp perpendicular to γ . The Jacobi equation for W gives

$$(a) \quad 0 = \frac{D^2W}{dt^2} + R(W, V)V = f''V + \frac{D^2W^\perp}{dt^2} + R(W^\perp, V)V.$$

Now

$$0 = \langle W^\perp, V \rangle \implies 0 = \left\langle \frac{DW^\perp}{dt}, V \right\rangle \implies 0 = \left\langle \frac{D^2W^\perp}{dt^2}, V \right\rangle$$

and we also have

$$0 = \langle R(W^\perp, V)V, V \rangle.$$

So (a) implies that $f'' = 0$, and therefore that W^\perp is a Jacobi field. Uniqueness is obvious.

(3) Write $W = fV + W^\perp$ as in (2). Then the linear function f must satisfy $f(a) = f(b) = 0$. So $f = 0$. ♦

Proposition 3 shows that for the purposes of investigating conjugate values, we need consider only perpendicular Jacobi fields. In particular, when M is a surface, and Y is a unit normal vector field along the geodesic $\gamma: [a, b] \rightarrow M$, any normal vector field W can be written uniquely as $W = gY$. We have $DY/dt = 0$, since γ is a geodesic and Y makes a constant angle with the parallel vector field $d\gamma/dt$. So the Jacobi equation for W becomes

$$g''(t)Y(t) + g(t)R(Y(t), V(t))V(t) = 0,$$

which is equivalent to

$$g''(t) + g(t)\langle R(Y(t), V(t))V(t), Y(t) \rangle = 0,$$

since we obtain $0 = 0$ when we take the inner product of the original equation with V . When the tangent vector $V = d\gamma/dt$ has length 1, we can write our equation as

$$g''(t) + K(\gamma(t)) \cdot g(t) = 0,$$

where K is the Gaussian curvature; this is the classical “Jacobi equation” for M .

The next theorem, basically due to Jacobi, gives a geometric way of obtaining Jacobi fields.

4. PROPOSITION. Let $\gamma: [a, b] \rightarrow M$ be a geodesic and let $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ be a variation of γ through geodesics, so that each $\bar{\alpha}(u): [a, b] \rightarrow M$ is also a geodesic. Then the variation vector field $W(t) = \partial\alpha/\partial u(0, t)$ is a Jacobi field along γ .

PROOF. Since α is a variation of γ through geodesics, we have

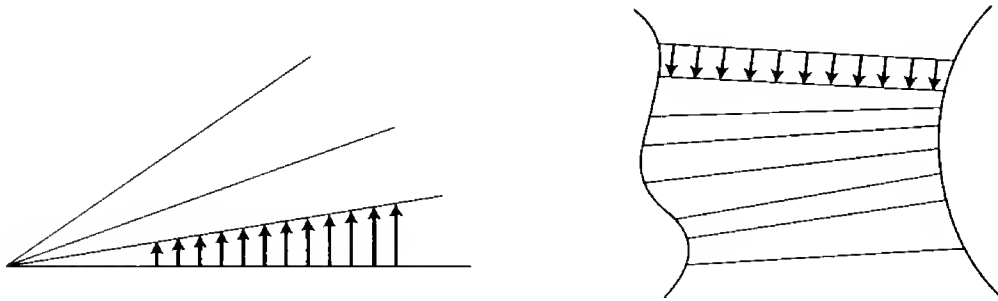
$$\frac{D}{\partial t} \frac{\partial \alpha}{\partial t} = 0.$$

Therefore

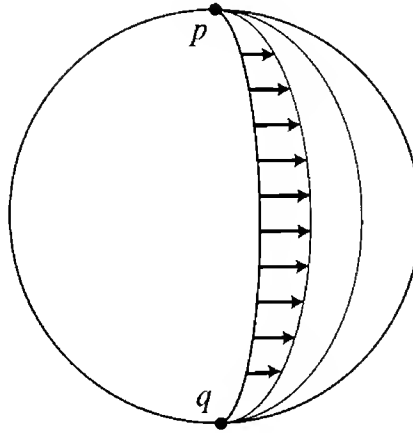
$$\begin{aligned} 0 &= \frac{D}{\partial u} \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} = \frac{D}{\partial t} \frac{D}{\partial u} \frac{\partial \alpha}{\partial t} + R\left(\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right) \frac{\partial \alpha}{\partial t} && \text{by Proposition II.6-10} \\ &= \frac{D^2}{\partial t^2} \frac{\partial \alpha}{\partial u} + R\left(\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right) \frac{\partial \alpha}{\partial t} && \text{by Proposition II.6-9,} \end{aligned}$$

which shows that $\partial\alpha/\partial u$ is a Jacobi field. ♦

Thus one way of obtaining Jacobi fields is to move geodesics around. In



particular, if γ is a great semi-circle on S^n , joining two antipodal points p and q , then a rotation of S^n keeping p and q fixed yields a variation vector field along γ which is a Jacobi field vanishing at p and q . Since we can rotate in $n - 1$ different directions, the points p and q have multiplicity $n - 1$, the theoretical maximum.



5. PROPOSITION. Every Jacobi field along a geodesic $\gamma: [a, b] \rightarrow M$ is the variation vector field of a variation of γ through geodesics.

PROOF. First suppose that γ lies completely inside an open set $U \subset M$ such that any two points $p, q \in U$ are joined by a unique geodesic in U , depending smoothly on p and q , of length $d(p, q)$. Given two vectors $W(a) \in M_{\gamma(a)}$ and $W(b) \in M_{\gamma(b)}$, choose curves $c_a, c_b: (-\varepsilon, \varepsilon) \rightarrow U$ such that

$$\begin{aligned} c_a(0) &= \gamma(a) & c_b(0) &= \gamma(b) \\ c_a'(0) &= W_a & c_b'(0) &= W_b. \end{aligned}$$

Define $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ by letting $\bar{\alpha}: [a, b] \rightarrow M$ be the unique geodesic in U from $c_a(u)$ to $c_b(u)$ of length $d(c_a(u), c_b(u))$. Then $W(t) = \partial\alpha/\partial u(0, t)$ is a Jacobi field along γ , by Proposition 4. To show that all Jacobi fields arise in this way, simply consider the map

$$\Phi: \{\text{Jacobi fields along } \gamma\} \rightarrow M_{\gamma(a)} \oplus M_{\gamma(b)}$$

given by

$$W \mapsto (W(a), W(b)).$$

We have just shown that Φ is onto. Since the domain and range of Φ both have dimension $2n$, the linear map Φ must also be one-one. Thus W is determined by $W(a), W(b)$; this shows that when the above construction is applied to $W(a)$ and $W(b)$, the resulting Jacobi field $\partial\alpha/\partial u(0, t)$, obtained by a variation through geodesics, is precisely the given Jacobi field W .

For a general geodesic γ , we note that for sufficiently small δ , the restricted geodesic $\gamma|_{[a, a+\delta]}$ will lie in an appropriate set U , by Theorem I.9-14. This gives us a variation through geodesics $\alpha: (-\varepsilon, \varepsilon) \times [a, a+\delta]$ with $\partial\alpha/\partial u(0, t)$ equal to the given Jacobi field $W(t)$ for $t \in [a, a+\delta]$. Using compactness of $[a, b]$, it is easy to see that if ε is made sufficiently small, then each geodesic $\bar{\alpha}(u)$ can be extended to a geodesic $\bar{\bar{\alpha}}(u): [a, b] \rightarrow M$. Then $(u, t) \mapsto \bar{\bar{\alpha}}(u)(t)$ is the required variation through geodesics. ♦

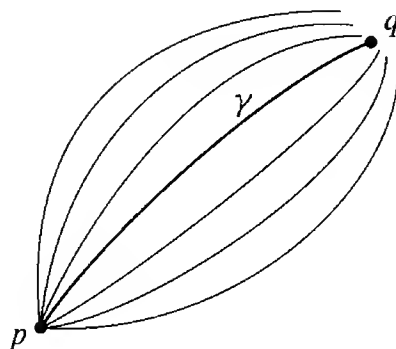
An examination of the proof of Proposition 5 shows that if W is a Jacobi field along a geodesic $\gamma: [a, b] \rightarrow M$ with $W(a) = 0$, then we can even find a variation $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ of γ through geodesics such that

$$\begin{aligned} \frac{\partial \alpha}{\partial u}(0, t) &= W(t) \\ \alpha(u, a) &= \gamma(a) \quad \text{for all } u \in (-\varepsilon, \varepsilon). \end{aligned}$$

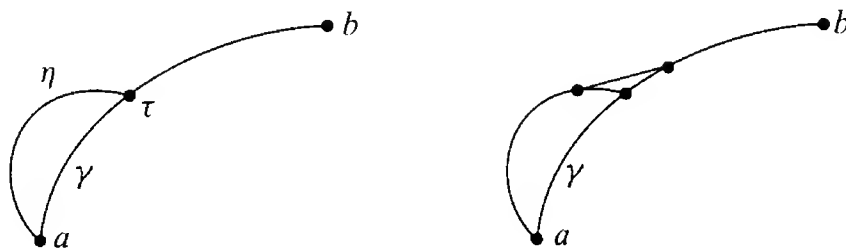
However, if $W(b) = 0$ for some other point b , we may not be able to choose α so that we also have $\alpha(u, b) = \gamma(b)$ for all u ; we will merely have this condition “up to first order”, that is,

$$\frac{\partial \alpha}{\partial u}(0, b) = 0.$$

Thus a conjugate point of $p = \gamma(a)$ is a place where some 1-parameter family of geodesics starting from p “nearly” intersect. This description of conjugate



points shows why they should play such an important role in the study of local minima for length, for it is easy to give an intuitive argument to prove that a geodesic $\gamma: [a, b] \rightarrow M$ *cannot* locally minimize length if there is some $\tau \in (a, b)$ conjugate to a . In fact, suppose we have another geodesic η from $\gamma(a)$ to $\gamma(\tau)$ with nearly the same length as $\gamma|_{[a, \tau]}$. Then γ has nearly the same length as η followed by $\gamma|_{[\tau, b]}$. But this compound curve has a corner, and can clearly be



made shorter by replacing the corner with a minimal geodesic. Therefore, γ is *not* a curve of minimal length. This reasoning turns out to be perfectly valid, provided that one works infinitesimally:

6. **THEOREM.** Let $\gamma: [a, b] \rightarrow M$ be a geodesic, and suppose that there is a number $\tau \in (a, b)$ which is conjugate to a along γ . Then there is some $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$. Consequently, γ is *not* a local minimum for E .

PROOF. Since τ is conjugate to a along γ , there is a non-zero Jacobi field J along γ such that $J(a) = J(\tau) = 0$. Let \tilde{J} be the vector field along γ with

$$\begin{aligned}\tilde{J}(t) &= J(t) & a \leq t \leq \tau \\ \tilde{J}(t) &= 0 & \tau \leq t \leq b.\end{aligned}$$

Notice that the discontinuity of $D\tilde{J}/dt$ at τ is

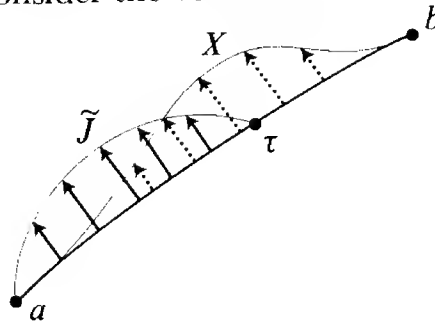
$$\Delta_\tau \frac{D\tilde{J}}{dt} = \frac{DJ}{dt}(\tau) \neq 0,$$

the inequality following from the fact that $J(\tau) = 0$, but J is non-zero. Choose a vector field X along γ which vanishes at a and b and which satisfies

$$(1) \quad \langle X(\tau), \Delta_\tau D\tilde{J}/dt \rangle = 1.$$

Now let c be a small number and consider the vector field

$$W = \frac{1}{c}\tilde{J} - cX.$$



We have

$$E_{**}(W, W) = \frac{1}{c^2} E_{**}(\tilde{J}, \tilde{J}) - 2E_{**}(\tilde{J}, X) + c^2 E_{**}(X, X).$$

Using the second variation formula, this becomes

$$\begin{aligned}E_{**}(W, W) &= 0 - 2\langle X(\tau), \Delta_\tau D\tilde{J}/dt \rangle + c^2 E_{**}(X, X) \\ &= -2 + c^2 E_{**}(X, X) \quad \text{by (1).}\end{aligned}$$

For sufficiently small c this is negative, which proves the first part of the theorem.

We have really already observed that the first part of the theorem implies the second, but we repeat the reasoning here. Suppose we have $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$. Consider the variations

$$\begin{aligned}\alpha(u, t) &= \exp uW(t) \\ \beta(u_1, u_2, t) &= \alpha(u_1 + u_2, t) = \exp(u_1 + u_2)W(t).\end{aligned}$$

Then

$$\begin{aligned} \left. \frac{\partial^2 E(\bar{\alpha}(u))}{\partial u^2} \right|_{u=0} &= \left. \frac{\partial^2 E(\bar{\beta}(u))}{\partial u_1 \partial u_2} \right|_{(u_1, u_2)=(0,0)} \\ &= E_{**}(W, W) < 0. \end{aligned}$$

So $u \mapsto E(\bar{\alpha}(u))$ has a strict relative maximum at $u = 0$. Therefore γ is not a relative minimum for E . ♦

Notice that the first part of this proof makes crucial use of the discontinuity of DW/dt , which is closely related to the kink in the “intuitive proof”. (Once we have obtained this W , however, we can always smooth it out to obtain an everywhere C^∞ vector field W with $E_{**}(W, W) < 0$.)

Our next hope is that a geodesic *does* minimize length among nearby paths if there are no conjugate points. In order to consider this case, we first need a result which contains essentially the same information as Propositions 4 and 5, but in a form that is much easier to use; for simplicity, we state it for a geodesic defined on $[0, 1]$.

7. THEOREM. Let $\gamma: [0, 1] \rightarrow M$ be a geodesic with $\gamma(0) = p \in M$ and $\gamma'(0) = v \in M_p$, so that γ can be described as $t \mapsto \exp tv$ for the map

$$\exp = \exp_p: M_p \rightarrow M.$$

Then 0 and 1 are conjugate values for γ if and only if v is a critical point of \exp .

PROOF. Suppose that v is a critical point for \exp . Then $\exp_*(X) = 0$ for some non-zero $X \in (M_p)_v =$ the tangent space of M_p at v . Let c be a path in M_p with $c(0) = v$ and $c'(0) = X$, and define

$$\alpha(u, t) = \exp tc(u) \quad 0 \leq t \leq 1.$$

Then α is a variation of γ through geodesics, so the vector field

$$W(t) = \left. \frac{\partial}{\partial u} \right|_{u=0} \exp tc(u)$$

is a Jacobi field along γ . We clearly have $W(0) = 0$, and also

$$\begin{aligned} W(1) &= \left. \frac{\partial}{\partial u} \right|_{u=0} \exp c(u) = \exp_* c'(0) \\ &= \exp_* X = 0. \end{aligned}$$

Moreover,

$$\begin{aligned}
 \frac{DW}{dt}(0) &= \frac{D}{\partial t} \Big|_{t=0} \frac{\partial}{\partial u} \Big|_{u=0} \exp tc(u) \\
 &= \frac{D}{\partial u} \Big|_{u=0} \frac{\partial}{\partial t} \Big|_{t=0} \exp tc(u) \quad \text{by Proposition II.6-9} \\
 &= \frac{D}{\partial u} \Big|_{u=0} c(u);
 \end{aligned}$$

this last expression is the covariant derivative of the vector field $u \mapsto c(u)$ along the constant curve $u \mapsto p$. Hence

$$\frac{DW}{dt}(0) = c'(0) = X \neq 0.$$

In particular, W is not identically 0, which shows that 0 and 1 are conjugate values for γ .

Now suppose that v is not a critical point for \exp . If $X_1, \dots, X_n \in (M_p)_v$ are n linearly independent vectors, then $\exp_*(X_1), \dots, \exp_*(X_n) \in M_{\gamma(1)}$ are also linearly independent. Choose paths c_1, \dots, c_n in M_p with $c_i(0) = v$ and $c_i'(0) = X_i$, and consider the variations

$$\alpha_i(u, t) = \exp tc_i(u),$$

with variation vector fields W_i . Then the W_i are Jacobi fields along γ which vanish at 0. Moreover, the $W_i(1) = \exp_*(X_i)$ are independent, so no non-trivial linear combination of the W_i can vanish at 1. Since the vector space of Jacobi fields along γ which vanish at 0 has dimension exactly n , it follows that no non-zero Jacobi field along γ vanishes at 0 and also at 1. ♦

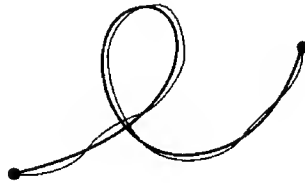
Since the points in M_p where \exp_* is zero form a closed set, Theorem 7 shows that the numbers τ conjugate to 0 along a geodesic $\gamma: [0, \infty) \rightarrow M$ also form a closed set. In particular, if there is any such τ , then there is a *first* τ conjugate to 0. Actually, much more is true, for the set of τ conjugate to 0 consists only of isolated points, so there are only finitely many τ conjugate to 0 in any interval $[0, b]$. We will not prove this here, but it is included in another result which we will state later on.

It is now a simple matter to prove the local length-minimizing property of a geodesic $\gamma: [a, b] \rightarrow M$ satisfying the condition that no number $\tau \in (a, b]$ is a conjugate value of a along γ . For simplicity, we will call such a γ a geodesic “without conjugate points”.

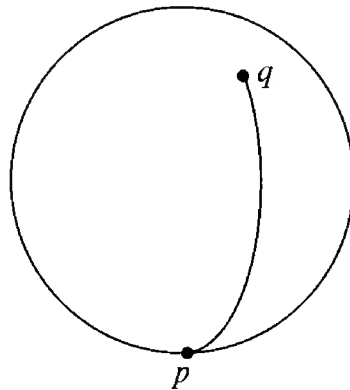
8. THEOREM. Let $\gamma: [a, b] \rightarrow M$ be a one-one geodesic with no conjugate points. Then γ has strictly smaller length than all sufficiently nearby paths between $p = \gamma(a)$ and $q = \gamma(b)$ (except for those which are merely reparameterizations of γ).

PROOF. By reparameterizing, we may assume that $[a, b] = [0, 1]$. If $v = \gamma'(0)$, then by Theorem 7 the map $\exp = \exp_p: M_p \rightarrow M$ is regular on the set $\{tv : 0 \leq t \leq 1\} \subset M_p$. By Lemma I.9-19 there is an open set $U \supset L$ on which \exp is a diffeomorphism. The result then follows from Problem I.9-29. ♦

Remark: Theorem 8 clearly remains true even for geodesics γ with self-intersections, provided that “nearby” paths refer to paths c with $c(t)$ close to $\gamma(t)$ for all t .



Let us test out Theorems 6 and 8 on the 2-sphere $S^2(r)$ of radius r , with $\gamma: [0, L] \rightarrow S^2(r)$ a portion of a great circle starting from a point p . We take γ



to be parameterized by arclength, so that $V = d\gamma/dt$ has length 1. Proposition 3(3) shows that in order to investigate conjugate points along γ , it suffices to consider Jacobi fields which are perpendicular to γ . If Y is a unit normal vector field along γ , then the Jacobi equation for $W = gY$ is (compare page 211)

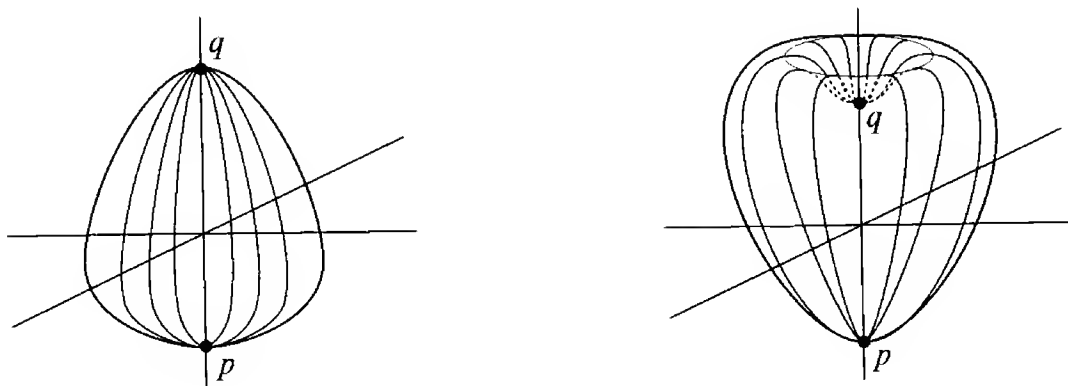
$$g''(t) + \frac{1}{r^2}g(t) = 0.$$

The solutions vanishing at $t = 0$ are all multiples of

$$g(t) = \sin \frac{t}{r},$$

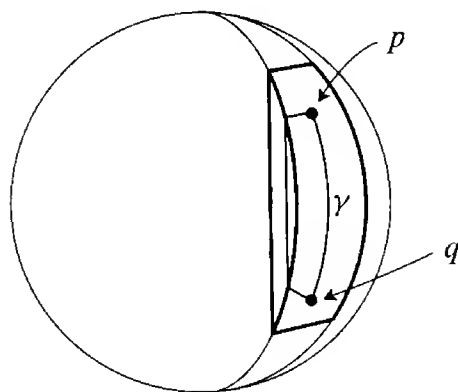
which has its first positive 0 at πr . So if $L > \pi r$, then γ contains a conjugate point, and Theorem 6 shows that γ does not locally minimize length. This is easy to see from the picture; in fact, in this case the intuitive proof of Theorem 6 works exactly. If $L < \pi r$, then Theorem 8 shows that γ does locally minimize length.

We have exactly the same situation for any compact surface of revolution M , when we take p to be one of the points where M intersects the z -axis I_z . The



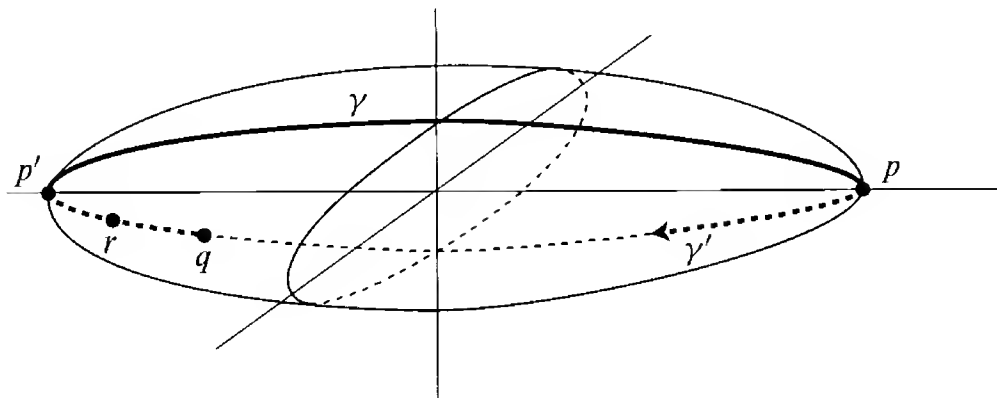
geodesics through p are the meridians, and it is clear, just by looking at the picture, that the only point conjugate to p along any geodesic is the other point q of $M \cap I_z$. Geodesics which do not reach q are local strict minima for length, and geodesics which extend past q are not local minima.

In this example it is clear that a geodesic γ which does not reach q is actually a minimum among all paths. [Proof: A minimum path between p and the other end of γ exists, since M is complete, and this path must be a geodesic; we know what all geodesics through p are, and γ is clearly the shortest.] However, it is easy to concoct examples where the non-existence of conjugate points implies only that γ is a *local* minimum for E . For example, we can round off the edges of the surface shown below (the boundary of part of a spherical wedge). Since



the surface is a sphere in a neighborhood of γ , it is still the case that no two points of γ are conjugate, and Theorem 8 still applies. On the other hand, there is clearly a shorter path between p and q if the wedge is thin enough.

A little more interesting situation arises for an ellipsoid. For the geodesic γ shown below, the first point q conjugate to p along γ occurs past the point p'

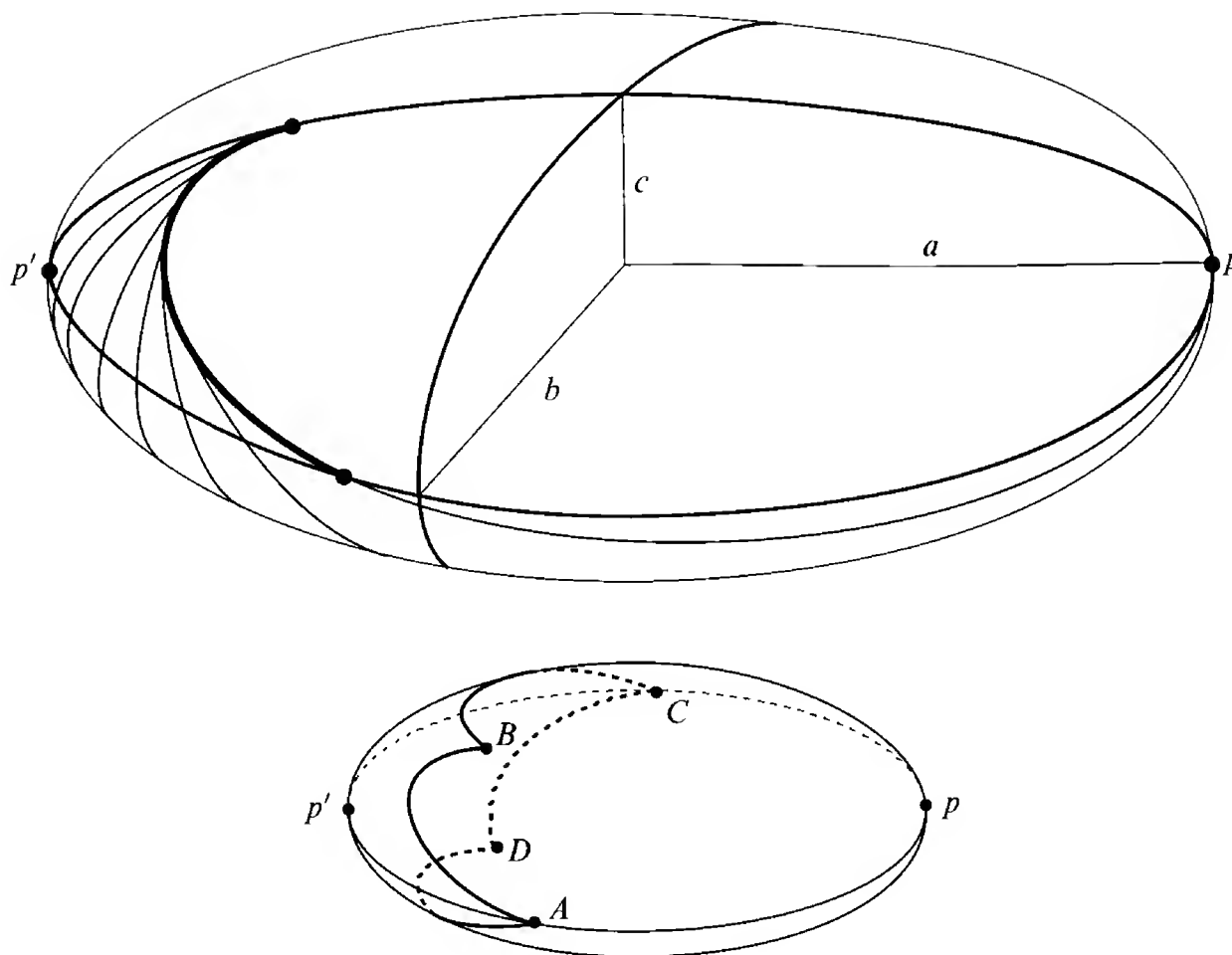


on the opposite end of the axis. (To establish this fact one has to examine the Jacobi equations for γ with some care.) If r is a point between p' and q , then the portion of γ between p and r is a local minimum for E , but clearly not a global minimum, since the extension γ' of γ in the other direction past p has shorter length from p to r (on the other hand, γ' is the only other geodesic from p to r which is shorter than γ).

Notice that Theorems 6 and 8 do not cover the case where b is the only point in $(a, b]$ which is conjugate to a . This is the borderline case for which no conclusions can be drawn. It may happen, first of all, that γ is a local minimum for length, but not a strict local minimum. This is illustrated, of course, by taking γ to be half of a great circle on the sphere S^2 . Now consider an ellipsoid, with three unequal axes $a > b > c$, and let p be a point at one end of the largest axis. The figure on the opposite page shows the conjugate points of the geodesics starting from p (it is the envelope of these geodesics—compare the Addendum to Chapter 3); this set is a curve with four cusps. The geodesic from p to A is a strict global minimum, while the geodesic going from p to p' and then on to B is a strict local minimum.

The next result complements Theorem 8 so that it appears to parallel Theorem 6 more closely.

9. PROPOSITION. Let $\gamma: [a, b] \rightarrow M$ be a geodesic without conjugate points. Then $E_{**}(W, W) > 0$ for every non-zero $W \in \Omega_\gamma$.



PROOF. Theorem 8 (and the Remark following it) implies that $E_{**}(W, W) \geq 0$. For if $E_{**}(W, W) < 0$, then γ would not be a local minimum for E (by the argument in the proof of Theorem 6).

Now suppose we had $E_{**}(W_1, W_1) = 0$ for some non-zero $W_1 \in \Omega_\gamma$. Then for any $W_2 \in \Omega_\gamma$ we would have

$$\begin{aligned} 0 &\leq E_{**}(W_1 + cW_2, W_1 + cW_2) \\ &= 0 + 2c E_{**}(W_1, W_2) + c^2 E_{**}(W_2, W_2). \end{aligned}$$

Since this is supposed to be true for all c , it is clear that we would have to have $E_{**}(W_1, W_2) = 0$. Thus W_1 would be a Jacobi field, contradicting the fact that b is not a conjugate value of a . ♦

More generally, we have the following result, which plays an important role later on.

10. COROLLARY. Let $\gamma: [a, b] \rightarrow M$ be a geodesic without conjugate points, W a Jacobi field along γ , and X a piecewise C^∞ vector field along γ

with

$$X(a) = W(a), \quad X(b) = W(b).$$

Then

$$E_{**}(X, X) \geq E_{**}(W, W),$$

and equality holds only when $X = W$.

PROOF. The second variation formula shows that for all piecewise C^∞ vector fields W_2 along γ we have

$$(1) \quad E_{**}(W, W_2) = \left\langle W_2, \frac{DW}{dt} \right\rangle \Big|_a^b = \left\langle W_2(b), \frac{DW}{dt}(b) \right\rangle - \left\langle W_2(a), \frac{DW}{dt}(a) \right\rangle.$$

Moreover, since $X - W \in \Omega_\gamma$, Proposition 9 shows that

$$\begin{aligned} 0 &\leq E_{**}(X - W, X - W) \\ &= E_{**}(X, X) + E_{**}(W, W) - 2E_{**}(W, X) \\ &= E_{**}(X, X) + \left\langle W, \frac{DW}{dt} \right\rangle \Big|_a^b - 2 \left\langle X, \frac{DW}{dt} \right\rangle \Big|_a^b \quad \text{by (1)} \\ &= E_{**}(X, X) - \left\langle W, \frac{DW}{dt} \right\rangle \Big|_a^b \quad \text{since } W = X \text{ at } a \text{ and } b \\ &= E_{**}(X, X) - E_{**}(W, W) \quad \text{by (1) again.} \end{aligned}$$

Moreover, it is clear that equality holds only if $X - W = 0$. ♦

Theorem 6 and Proposition 9 show that for a geodesic $\gamma: [a, b] \rightarrow M$, the existence of conjugate points is practically equivalent to the existence of vector fields $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$:

- (A) If there is some $\tau \in (a, b)$ conjugate to a , then there is some $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$ (Theorem 6);
- (B) If there is some $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$, then there is some $\tau \in (a, b]$ conjugate to a (Proposition 9).

We will see later that it can be very convenient to replace questions about conjugate points by questions about vector fields $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$. Actually, the situation is even better than we have indicated, because statement (B) can be strengthened: if $E_{**}(W, W) < 0$ for some $W \in \Omega_\gamma$, then there is $\tau \in (a, b)$ conjugate to a . In fact, there is a far-reaching generalization of

these results. We say that E_{**} is **negative definite** on a subspace $\mathcal{V} \subset \Omega_\gamma$ if $E_{**}(W, W) < 0$ for all non-zero $W \in \mathcal{V}$, and we define the **index** of E_{**} to be the largest dimension of any subspace $\mathcal{V} \subset \Omega_\gamma$ on which E_{**} is negative definite (compare page 3). Then we have the celebrated

MORSE INDEX THEOREM. The index of E_{**} for a geodesic $\gamma: [a, b] \rightarrow M$ is the number of $\tau \in (a, b)$ which are conjugate to a , each conjugate value being counted with its multiplicity. This index is always finite.

In terms of the index of E_{**} , our Theorem 6 can be reformulated as follows: if the number of conjugate values is ≥ 1 , then the index is ≥ 1 . For the Morse Index Theorem we need the more general assertion, that the index of $(E_a^t)_{**}$ increases by at least ν as t passes a conjugate value τ with multiplicity ν . This is the *only* point in the proof that does not involve simple general considerations, and it may be handled by essentially the same trick which was used in the present proof of Theorem 6. I hope that by clearing this path right up to the proof of the Index Theorem, I may have enticed you into reading the proof in Milnor {2}, which also describes some of the beautiful applications of these differential geometric ideas to topology.

In order to obtain interesting differential geometric consequences of our foundational results, we need to find hypotheses which imply something about the solutions of Jacobi equations. These hypotheses usually involve the sectional curvature $K(P)$ of 2-dimensional subspaces $P \subset M_p$; recall that $K(P) = \langle R(X, Y)Y, X \rangle$ for orthonormal $X, Y \in P$. Clearly all sectional curvatures of M are ≤ 0 if and only if $\langle R(X, Y)Y, X \rangle \leq 0$ for all pairs X, Y of vectors at the same point of M .

11. PROPOSITION. If all sectional curvatures of M are ≤ 0 , then no two points of M are conjugate along any geodesic.

PROOF. If γ is a geodesic with velocity vector field $V = d\gamma/dt$, and W is a Jacobi field along γ , then

$$\frac{D^2 W}{dt^2} + R(W, V)V = 0,$$

so

$$\left\langle \frac{D^2 W}{dt^2}, W \right\rangle = -\langle R(W, V)V, W \rangle \geq 0.$$

Therefore

$$\frac{d}{dt} \left\langle \frac{DW}{dt}, W \right\rangle = \left\langle \frac{D^2 W}{dt^2}, W \right\rangle + \left\langle \frac{DW}{dt}, \frac{DW}{dt} \right\rangle \geq 0,$$

which means that $\langle DW/dt, W \rangle$ is increasing.

Now if W vanishes at two points, t_0 and t_1 , then $\langle DW/dt, W \rangle$ vanishes at t_0 and t_1 , so $\langle DW/dt, W \rangle$ must be 0 on the interval $[t_0, t_1]$. This clearly implies that DW/dt also vanishes at t_0 . Hence $W = 0$. ♦

Although Proposition 11 shows that all geodesic segments on M are local minima for length, this does not mean that they are necessarily global minima. In fact, if we consider a compact surface M with everywhere negative curvature (Chapter 6, Addendum 1), it is clear that no geodesic $\gamma: \mathbb{R} \rightarrow M$ can be a global minimum for length on arbitrarily large segments.

The most interesting consequence of Proposition 11 is obtained by combining it with the following general result.

12. THEOREM. Let M be a complete, connected, n -dimensional Riemannian manifold, and let p be a point of M such that no point of M is conjugate to p along any geodesic. Then $\exp = \exp_p: M_p \rightarrow M$ is a covering map. In particular, if M is simply-connected, then M is diffeomorphic to \mathbb{R}^n .

13. COROLLARY (HADAMARD-CARTAN). A complete, simply-connected, n -dimensional Riemannian manifold with all sectional curvatures ≤ 0 is diffeomorphic to \mathbb{R}^n .

PROOF. The Corollary follows immediately from the Theorem and Proposition 11. To prove the Theorem, let $\langle \cdot, \cdot \rangle$ be the Riemannian metric on M , and consider the tensor $\exp^*\langle \cdot, \cdot \rangle$ on M_p . Since there are no points conjugate to p , the map \exp_* is always one-one, so $\exp^*\langle \cdot, \cdot \rangle$ is a Riemannian metric on M_p . We claim that M_p is complete in the metric $\exp^*\langle \cdot, \cdot \rangle$. To prove this, we just note that the straight lines through 0 in M_p are clearly geodesics for the metric $\exp^*\langle \cdot, \cdot \rangle$, since their images under the local isometry $\exp: M_p \rightarrow M$ are geodesics in M . Since all geodesics through $0 \in M_p$ can be defined for all t , it follows from Problem I.9-43 that M_p is complete. The Theorem then follows from

14. LEMMA. Let M and N be connected Riemannian manifolds with M complete, and let $\phi: M \rightarrow N$ be a local isometry. Then N is complete and ϕ is a covering map onto N .

PROOF. Let $p_0 \in M$. Given a geodesic $\gamma: (-\varepsilon, \varepsilon) \rightarrow N$ with $\gamma(0) = \phi(p_0)$, let c be the geodesic in M with $c(0) = p_0$ and $\phi_*c'(0) = \gamma'(0)$. Then $\gamma = \phi \circ c$,

since ϕ is a local isometry. Since c can be defined on all of \mathbb{R} , we can extend γ to all of \mathbb{R} as $\phi \circ c$. Thus N is complete, by Problem I.9-43.

To prove that ϕ is onto N it suffices to prove that $\phi(M)$ is closed (for $\phi(M)$ is open, since ϕ is everywhere regular). Let $q \in \overline{\phi(M)}$, and let V be a convex neighborhood of 0 in N_q on which \exp_q is a diffeomorphism. There is a point $q' \in \exp_q(V)$ of the form $q' = \phi(p')$ for $p' \in M$. Let γ be the geodesic in $\exp_q(V)$ with $\gamma(0) = q'$ and $\gamma(1) = q$. Consider the geodesic c in M with $c(0) = p'$ and $\phi_*c'(0) = \gamma'(0)$. Then $\gamma = \phi \circ c$, as before. The point $p = c(1)$ is defined and $\phi(p) = \phi(c(1)) = \gamma(1) = q$. Thus $\overline{\phi(M)} \subset \phi(M)$, so $\phi(M)$ is closed. Hence ϕ is onto N .

The proof that ϕ is a covering map will be similar to the proof that appears on pp. III.258–259. For fixed $q \in N$, let

$$V = \{Y \in N_q : \|Y\| < 2\varepsilon\} \subset N_q$$

be a neighborhood of 0 in N_q on which \exp_q is a diffeomorphism. Suppose that $p \in \phi^{-1}(q)$. Consider the map

$$\begin{aligned} f &= \exp_p \circ \phi_{p*}^{-1} \circ (\exp_q(V))^{-1}, \\ f: \exp_q(V) &\rightarrow \exp_p(\{X \in M_p : \|X\| < 2\varepsilon\}) \subset M; \end{aligned}$$

this map is defined since M is complete. It is easy to see that

$$\phi: \exp_p(\{X \in M_p : \|X\| < 2\varepsilon\}) \rightarrow \exp_q(V)$$

is a diffeomorphism with inverse f . Now let

$$W = \exp_q(\{Y \in N_q : \|Y\| < \varepsilon\}) \subset N,$$

and for each $p \in M$, let

$$W_p = \exp_p(\{X \in M_p : \|X\| < \varepsilon\}) \subset M.$$

We claim that

$$\phi^{-1}(W) = \bigcup_{p \in \phi^{-1}(q)} W_p.$$

In fact, given $p' \in \phi^{-1}(W)$, let γ be the geodesic in W of length $d(\phi(p'), q)$ with $\gamma(0) = \phi(p')$ and $\gamma(1) = q$. Let c be the geodesic with $c(0) = p'$ and $\phi_*c'(0) = \gamma'(0)$. Then c is defined on $[0, 1]$, since M is complete, and $\phi \circ c = \gamma$ on $[0, 1]$. In particular, $p = c(1) \in \phi^{-1}(q)$, and it is easy to see that all points of $c([0, 1])$ are in W_p . Thus $p' = c(0) \in W_p$.

To complete the proof we just have to show that the W_p are disjoint. Now if $W_{p_1} \cap W_{p_2} \neq \emptyset$, then we clearly have

$$p_2 \in \exp_{p_1}(\{X \in M_{p_1} : \|X\| < 2\varepsilon\}).$$

But we know that ϕ is a diffeomorphism on this set. Since $\phi(p_1) = \phi(p_2)$, it follows that $p_1 = p_2$. ♦

Proposition 11 is but a special case of more general results involving manifolds whose sectional curvatures satisfy certain inequalities. These results all follow from one theorem, but the mere statement of this theorem tends to overwhelm one with its complexity. So we will approach it rather gingerly by first proving special cases, all of which represent important steps in the historical evolution of the final result.

The first theorem of this type depends on a surprisingly simple proposition about second order differential equations. Remember that a solution ϕ of such an equation is determined by $\phi(a)$ and $\phi'(a)$. Consequently, a non-zero solution ϕ must have isolated zeros.

15. THEOREM (THE STURM COMPARISON THEOREM). Let f and h be two continuous functions satisfying $f(t) \leq h(t)$ for all t in an interval I , and let ϕ and η be two functions satisfying the differential equations

$$\begin{aligned} (1) \quad & \phi'' + f\phi = 0 \\ (2) \quad & \eta'' + h\eta = 0 \end{aligned}$$

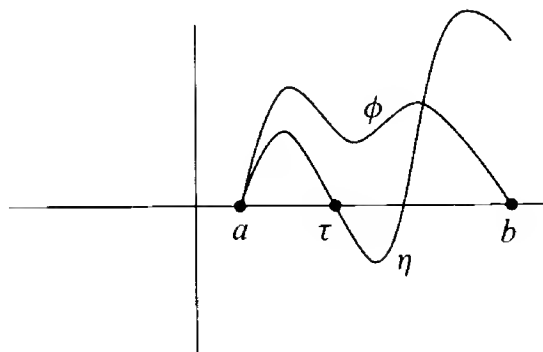
on I . Assume that ϕ is not the zero function, and let $a, b \in I$ be two consecutive zeros of ϕ .

(1) The function η must have a zero in (a, b) , unless $f = h$ everywhere on $[a, b]$ and η is a constant multiple of ϕ on $[a, b]$.

(2) Suppose that we have $\eta(a) = 0$, and also $\eta'(a) = \phi'(a) > 0$ [which can be achieved by choosing a suitable multiple of η , and changing ϕ to $-\phi$, if necessary]. If τ is the smallest zero of η in $(a, b]$, then

$$\phi(t) \geq \eta(t) \quad \text{for } a \leq t \leq \tau,$$

and equality holds for some t only if $f = h$ on $[a, t]$.



PROOF. Equations (1) and (2) give

$$(3) \quad \phi''\eta - \eta''\phi = (h - f)\phi\eta.$$

Suppose that η were nowhere zero on (a, b) . It is easy to see that there is no loss of generality in assuming that

$$(4) \quad \eta, \phi > 0 \quad \text{on } (a, b).$$

Then (3) gives

$$0 \leq \phi''\eta - \eta''\phi,$$

so

$$(5) \quad 0 \leq \int_a^b \phi''\eta - \eta''\phi = \int_a^b (\phi'\eta - \eta'\phi)' \\ = \phi'(b)\eta(b) - \phi'(a)\eta(a), \quad \text{since } \phi(a) = \phi(b) = 0.$$

On the other hand, (4) clearly implies that

$$(6) \quad \left\{ \begin{array}{l} \phi'(a) > 0, \phi'(b) < 0 \\ \eta(a), \eta(b) \geq 0 \end{array} \right\} \implies \phi'(b)\eta(b) - \phi'(a)\eta(a) \leq 0.$$

If $f \neq h$, then we actually have strict inequality in (5), which contradicts the second part of (6). This contradiction shows that η must have a zero on (a, b) .

If $f = h$ on $[a, b]$, then equality holds in (5), and the first part of (6) implies that we must have $\eta(a) = \eta(b) = 0$. Since ϕ and η then satisfy the same second order equation on $[a, b]$ and $\phi(a) = \eta(a)$, the solution η must be a constant multiple of ϕ on $[a, b]$.

Now suppose that $\eta(a) = 0$ and $\eta'(a) = \phi'(a) > 0$. If τ is the smallest zero of η in $(a, b]$, then $\phi, \eta > 0$ on (a, τ) , so (3) gives

$$0 \leq \phi''\eta - \eta''\phi = (\phi'\eta - \eta'\phi)' \quad \text{on } (a, \tau).$$

This implies that

$$0 \leq \phi'\eta - \eta'\phi \quad \text{on } (a, \tau),$$

since $[\phi'\eta - \eta'\phi](a) = 0$. Using positivity of η on (a, τ) again, this gives

$$(7) \quad 0 \leq \left(\frac{\phi}{\eta} \right)' \quad \text{on } (a, \tau).$$

But

$$\begin{aligned}\lim_{t \rightarrow a} \frac{\phi(t)}{\eta(t)} &= \lim_{t \rightarrow a} \frac{\phi'(t)}{\eta'(t)} && \text{by L'Hôpital's Rule} \\ &= 1, && \text{by assumption.}\end{aligned}$$

Therefore

$$\frac{\phi}{\eta} \geq 1 \quad \text{on } (a, \tau),$$

which is the desired inequality. The proof of the final statement is left to the reader. ♦

Remark 1: Since $\phi'(a)$ and $\eta'(a)$ exist, and $\eta'(a) \neq 0$, we really used only a trivial case of L'Hôpital's Rule; we could have simply written

$$\begin{aligned}\lim_{t \rightarrow a} \frac{\phi(t)}{\eta(t)} &= \lim_{t \rightarrow a} \frac{\phi(t) - \phi(a)}{\eta(t) - \eta(a)} = \lim_{t \rightarrow a} \frac{\frac{\phi(t) - \phi(a)}{t - a}}{\frac{\eta(t) - \eta(a)}{t - a}} \\ &= \frac{\phi'(a)}{\eta'(a)} = 1.\end{aligned}$$

Remark 2: In our applications, we will be interested only in the case where $\eta(a) = 0$. The reasoning for part (1) is then unnecessary, because part (2) shows that $\phi \geq \eta$ on any interval (a, τ) on which $\phi, \eta > 0$; this clearly implies that η vanishes somewhere on $(a, b]$. Moreover, if b were the first zero of η , then we would have $\phi(b) = \eta(b) = 0$, so we would have $f = h$ on $[a, b]$, by the final statement in part (2). Nevertheless, part (1) is still of interest; here is one consequence:

16. COROLLARY. If ϕ_1 and ϕ_2 are two linearly independent solutions of the equation

$$\phi'' + f\phi = 0,$$

then the zeros of ϕ_1 alternate with the zeros of ϕ_2 .

A particularly simple instance of Corollary 16 is provided by the equation $y'' + y/r^2 = 0$, where $r > 0$ is a constant. The solutions of this equation can all be written in the form $y(t) = b \sin(a + t/r)$. The zeros are always πr apart, so the zeros of two linearly independent solutions alternate with each other. This simple equation serves as a standard with which we can compare the Jacobi equation.

17. **THEOREM (BONNET).** Let M be a surface, and $\gamma: [0, L] \rightarrow M$ a geodesic parameterized by arclength. Let $r > 0$ be a constant.

(1) If $K(p) \leq 1/r^2$ for all $p = \gamma(t)$, and γ has length $L < \pi r$, then γ contains no conjugate points.

(2) If $K(p) \geq 1/r^2$ for all $p = \gamma(t)$, and γ has length $L > \pi r$, then there is a point $\tau \in (0, L)$ conjugate to 0, and therefore γ is not of minimal length.

(3) If M is connected and complete, and $K(p) \geq 1/r^2$ for all $p \in M$, then M is actually compact, with diameter $\leq \pi r$.

PROOF. (1) Let Y be a unit vector field along γ with $\langle V, Y \rangle = 0$, where V is the unit vector field $V = d\gamma/dt$. The Jacobi equation for the vector field ϕY is (compare page 211)

$$\phi''(t) + K(\gamma(t)) \cdot \phi(t) = 0.$$

The simpler equation

$$\eta''(t) + \frac{1}{r^2} \eta(t) = 0$$

has the solution $\eta(t) = \sin t/r$. Since $K(\gamma(t)) \leq 1/r^2$ by hypothesis, the Sturm comparison theorem shows that the first equation cannot have a solution ϕ vanishing at 0 and at $L < \pi r$, since η has no zero in $(0, L)$.

(2) Let Y be as in part (1), and consider a vector field ηY . The Jacobi equation for ηY is

$$\eta''(t) + K(\gamma(t)) \cdot \eta(t) = 0,$$

and the simpler equation

$$\phi''(t) + \frac{1}{r^2} \phi(t) = 0$$

has the solution $\phi(t) = \sin t/r$ which vanishes at 0 and at πr . Since $1/r^2 \leq K(\gamma(t))$, the comparison theorem shows that any Jacobi field ηY must have a zero on the open interval $(0, \pi r) \subset (0, L)$. So if we choose any non-zero Jacobi field ηY along γ with $\eta(0) = 0$, then this Jacobi field will also vanish at some $\tau \in (0, L)$; thus τ is conjugate to 0.

(3) Any two points $p, q \in M$ can be joined by a geodesic γ of minimal length (Theorem I.9-18). Then the length of γ must be $\leq \pi r$, by part (2). So M is bounded, with diameter $\leq \pi r$. Since closed bounded sets in a complete manifold are compact, it follows that M itself is compact. ♦

I do not know whether Sturm ever saw this beautiful application of his theorem (he died in 1855, the same year that Bonnet published the result), but in his lectures he is supposed to have referred to it as the theorem “whose name I have the honor to bear”.

Bonnet’s Theorem fairly cries out to be generalized to higher dimensional manifolds, but a direct approach leads us into difficulties. The single normal vector field Y along γ has to be replaced by $n - 1$ vector fields Y_1, \dots, Y_{n-1} . Even if we choose Y_1, \dots, Y_{n-1} to be parallel, everywhere orthonormal vector fields along γ , the Jacobi equation for $\sum_i \phi_i Y_i$ reads

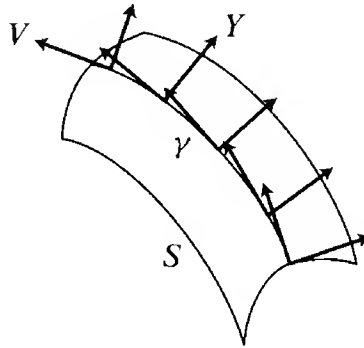
$$\sum_i \phi_i''(t) Y_i(t) + \sum_i \phi_i(t) \cdot R(Y_i(t), V(t)) V(t) = 0,$$

which is equivalent to a *system* of ordinary differential equations

$$\phi_j''(t) + \sum_i \phi_i(t) \langle R(Y_i(t), V(t)) V(t), Y_j(t) \rangle = 0,$$

and these equations do not even involve the sectional curvature directly. It is clear that we will have to approach this problem with a little more finesse.

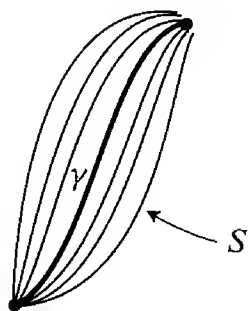
One way to extend the results of Bonnet’s theorem to higher dimensions is by an artful use of Synge’s inequality (Corollary 1-7). Suppose first that $K(P) \geq 1/r^2$ for all 2-dimensional $P \subset M_{\gamma(t)}$, and that $\gamma: [0, L] \rightarrow M$ has length $L > \pi r$. Let Y be a *parallel* vector field along γ which is everywhere perpendicular to the parallel vector field $V = d\gamma/dt$, and let $S \subset M$ be a surface containing γ whose tangent space at each point $\gamma(t)$ is spanned by $V(t)$ and $Y(t)$. Then



Synge’s inequality shows that the Gaussian curvature of S at $\gamma(t)$ is the same as the sectional curvature of M for the plane spanned by $V(t)$ and $Y(t)$. So the Gaussian curvature of S at $\gamma(t)$ is $\geq 1/r^2$. Bonnet’s theorem then shows that there is a point $\tau \in (0, L)$ conjugate to 0 along γ . Of course, this means that τ is a conjugate value for 0 *in the surface* S . To conclude that there is a

conjugate value in M itself, we must use the following indirect line of reasoning: Since $\tau \in (0, L)$ is a conjugate value for 0 in S , the geodesic γ is *not* a local minimum for length in S . Therefore it is certainly not a local minimum for length in M . Therefore, some $\sigma \in (0, L]$ must be a conjugate value for 0 in M . Applying this result to $\gamma|_{[0, L']}$ for $\pi r < L' < L$, we see that actually some $\sigma \in (0, L'] \subset (0, L)$ is conjugate to 0 along γ in M .

In the previous paragraph, we had to choose the surface S so as not to decrease K , and then we had to show that the choice of S made no difference for the final conclusion. If we instead try to analyze the case where $K(P) \leq 1/r^2$ and γ has length $L < \pi r$, then we certainly don't care whether K is decreased, but our choice of S will be much more dependent on the desired conclusion. Suppose, then, that γ contained a conjugate point $\tau \in (0, L]$. We might as well assume that L itself is conjugate to 0 along γ , since we can always work with $\gamma|_{[0, \tau]}$; for the same reason, we might as well assume that L is the smallest value conjugate to 0. Then there is a variation α of γ through geodesics in M , whose variation vector field $W(t) = \partial\alpha/\partial u(0, t)$ vanishes only at 0 and L . Consider the surface S formed by the image of α . Synge's inequality shows that the



Gaussian curvature of S is $\leq 1/r^2$ along γ , and then Bonnet's theorem shows that γ cannot contain a conjugate point on S . But γ clearly does contain a conjugate point on S , because the $\bar{\alpha}(u)$ are also geodesics on S , so the variation vector field of α is also a Jacobi field along γ on S . We seem to have obtained a contradiction, and thereby shown that γ cannot contain a conjugate point $\tau \in (0, L]$. The trouble with this argument is that only an excess of generosity could lead one to call S a surface, as the map α is definitely *not* an immersion at $(0, 0)$ or $(0, L)$. The idea of the proof is basically sound, however, and leads to the desired result if one reasons a little more carefully (Problem 2).

We have not bothered to bestow upon this reasoning the dignity which might accrue to it as the official proof of a theorem, because the results, and even better ones, can be obtained in a more systematic way. In fact, we shall present

two new methods of generalizing Bonnet's theorem. These two methods differ significantly in their basic philosophy, but they both depend on a certain construction, similar to one used above, which is best set forth in a separate Lemma. This very general sounding Lemma involves geodesics in two different manifolds, although in applications one of the two is always taken to be a sphere. (In the statement and proof of the Lemma, we will not use subscripts to distinguish the norms $\| \cdot \|$ and covariant derivatives D/dt in the two manifolds, since it should always be clear which manifold we are working in.)

18. LEMMA. Let M_1 and M_2 be two manifolds of the same dimension n , and let $\gamma_i: [a, b] \rightarrow M_i$ be arclength parameterized geodesics in these two manifolds. Then there is a vector space isomorphism

$$\begin{aligned} \Phi: \{\text{piecewise } C^\infty \text{ vector fields along } \gamma_1\} \\ \rightarrow \{\text{piecewise } C^\infty \text{ vector fields along } \gamma_2\} \end{aligned}$$

such that for all $t \in [a, b]$ we have

- (1) If $\frac{DX}{dt}$ is continuous at t , then $\frac{D\Phi(X)}{dt}$ is continuous at t ,
- (2) $\langle X(t), \gamma_1'(t) \rangle = \langle \Phi(X)(t), \gamma_2'(t) \rangle$,
- (3) $\|X(t)\| = \|\Phi(X)(t)\|$,
- (4) $\left\| \frac{DX}{dt}(t) \right\| = \left\| \frac{D\Phi(X)}{dt}(t) \right\|$,

it being understood that the last equation refers to left and right hand limits at discontinuity points.

PROOF. Pick some fixed $t_0 \in [a, b]$. Let $\phi: (M_1)_{\gamma_1(t_0)} \rightarrow (M_2)_{\gamma_2(t_0)}$ be any norm preserving isomorphism with $\phi(\gamma_1'(t_0)) = \gamma_2'(t_0)$. Then we can define

$$\phi_t: (M_1)_{\gamma_1(t)} \rightarrow (M_2)_{\gamma_2(t)}$$

by parallel translating a vector in $(M_1)_{\gamma_1(t)}$ along γ_1 to $\gamma_1(t_0)$, applying ϕ , and then parallel translating along γ_2 to $(M_2)_{\gamma_2(t)}$. We then define $\Phi(X)$ by

$$\Phi(X)(t) = \phi_t(X(t)).$$

We can also describe $\Phi(X)$ as follows. Let Y_1, \dots, Y_n be parallel, everywhere orthonormal vector fields along γ_1 with $Y_1(t_0) = \gamma_1'(t_0)$, and let Z_1, \dots, Z_n be

parallel, everywhere orthonormal vector fields along γ_2 with $Z_1(t_0) = \gamma_2'(t_0)$.
If

$$X(t) = \sum_{i=1}^n f_i(t) Y_i(t)$$

for certain functions $f_i: [a, b] \rightarrow \mathbb{R}$, then

$$\Phi(X)(t) = \sum_{i=1}^n f_i(t) Z_i(t).$$

This shows that $\Phi(X)$ is C^∞ everywhere that X is, and that

$$\langle X(t), \gamma_1'(t) \rangle = f_1(t) = \langle \Phi(X)(t), \gamma_2'(t) \rangle$$

$$\|X(t)\|^2 = \sum_{i=1}^n [f_i(t)]^2 = \|\Phi(X)(t)\|^2$$

$$\left\| \frac{DX}{dt}(t) \right\|^2 = \sum_{i=1}^n [f_i'(t)]^2 = \left\| \frac{D\Phi(X)}{dt}(t) \right\|^2. \quad \spadesuit$$

In our first generalization of Bonnet's Theorem, we will consider the index of a geodesic, instead of the number of conjugate points it contains. Recall (page 223) that the index of γ is > 0 if and only if there is some $W \in \Omega_\gamma$ with $E_{**}(W, W) < 0$.

19. THEOREM. Let M_1 and M_2 be two manifolds of the same dimension n , and let $\gamma_i: [a, b] \rightarrow M_i$ be geodesics parameterized by arclength. For each $t \in [a, b]$, suppose that for all 2-dimensional $P_i \subset (M_i)_{\gamma_i(t)}$, the curvatures K_i satisfy

$$K_1(P_1) \leq K_2(P_2).$$

Then we have

$$\text{index } \gamma_1 \leq \text{index } \gamma_2.$$

In particular, if $E_{**}(W_1, W_1) < 0$ for some $W_1 \in \Omega_{\gamma_1}$, then also $E_{**}(W_2, W_2) < 0$ for some $W_2 \in \Omega_{\gamma_2}$.

PROOF. Let W be a piecewise C^∞ vector field on γ_1 , and let Φ be the map in Lemma 18. The second variation formula shows that

$$\begin{aligned} (1) \quad E_{**}(W, W) = & - \int_a^b \langle R(W(t), V(t))V(t), W(t) \rangle dt \\ & - \int_a^b \left\langle W(t), \frac{D^2 W}{dt^2}(t) \right\rangle dt - \sum_{i=0}^N \left\langle W(t_i), \Delta_{t_i} \frac{DW}{dt} \right\rangle. \end{aligned}$$

Now we also have

$$\frac{d}{dt} \left\langle W(t), \frac{DW}{dt}(t) \right\rangle = \left\langle \frac{DW}{dt}(t), \frac{DW}{dt}(t) \right\rangle + \left\langle W(t), \frac{D^2W}{dt^2}(t) \right\rangle.$$

Integrating this equation between t_{i-1} and t_i for each i , and adding the results, we obtain

$$-\sum_{i=0}^N \left\langle W(t_i), \Delta_{t_i} \frac{DW}{dt_i} \right\rangle = \int_a^b \left\langle \frac{DW}{dt}(t), \frac{DW}{dt}(t) \right\rangle dt + \int_a^b \left\langle W(t), \frac{D^2W}{dt^2}(t) \right\rangle dt.$$

So equation (1) can be written

$$E_{**}(W, W) = \int_a^b \left\{ \left\langle \frac{DW}{dt}(t), \frac{DW}{dt}(t) \right\rangle - \langle R(W(t), V(t))V(t), W(t) \rangle \right\} dt.$$

From the properties of the map Φ , and the hypotheses on K , we see that

$$E_{**}(W, W) \geq E_{**}(\Phi(W), \Phi(W)).$$

So, if $\mathcal{V} \subset \Omega_{\gamma_1}$ is a subspace on which E_{**} is negative definite, then $\Phi(\mathcal{V}) \subset \Omega_{\gamma_2}$ is a subspace of the same dimension on which E_{**} is again negative definite. Thus the index of γ_2 is certainly at least as large as the index of γ_1 . ♦

20. COROLLARY (THE MORSE-SCHOENBERG COMPARISON THEOREM). Let M be a Riemannian manifold of dimension n , and let $\gamma: [0, L] \rightarrow M$ be a geodesic parameterized by arclength. Let $r > 0$ be a constant.

(1) If $K(P) \leq 1/r^2$ for all $P \in M_{\gamma(t)}$, and γ has length $L < \pi r$, then the index of γ is 0, and γ contains no conjugate point. [Note that Proposition 11 is a special case.]

(2) If $K(P) \geq 1/r^2$ for all $P \in M_{\gamma(t)}$, and γ has length $L > \pi r$, then there is a point $\tau \in (0, L)$ conjugate to 0, and γ is not of minimal length.

PROOF. (1) We apply the Theorem with $M_1 = M$ and $M_2 = n$ -sphere $S^n(r)$ of radius r , choosing γ_1 to be γ , and $\gamma_2: [0, L] \rightarrow S^n(r)$ to be any geodesic parameterized by arclength. We find that

$$\text{index } \gamma \leq \text{index } \gamma_2.$$

Now the index of γ_2 is certainly zero, since γ_2 contains no conjugate points, and Proposition 9 applies (all we really need is the fact that $E_{**}(W, W) \geq 0$

for $W \in \Omega_{\gamma_2}$, which follows from Theorem 8). Consequently, index $\gamma = 0$. Theorem 6 implies that no number $\tau \in (0, L)$ is conjugate to 0 along γ . We can also conclude that no number $\tau \in (0, L]$ is conjugate to 0 along γ , by extending γ to $\bar{\gamma}: [0, L'] \rightarrow M$ with $L < L' < \pi r$, and applying the result to $\bar{\gamma}$.

(2) We apply the Theorem with $M_1 = S^n(r)$ and $M_2 = M$, this time choosing γ_2 to be γ . We obtain

$$\text{index } \gamma_1 \leq \text{index } \gamma.$$

But the index of γ_1 is at least 1, since γ_1 contains a conjugate point, and Theorem 6 applies. Consequently, index $\gamma \geq 1$. This shows that γ does contain a conjugate point $\tau \in (0, L]$ (Proposition 9 or Theorem 8 again). Applying this result to $\gamma|_{[0, L']}$, with $\pi r < L' < L$, we see that γ contains a conjugate value $\tau \in (0, L)$. ♦

For the case where $K \geq 1/r^2$, we can obtain a stronger result, involving the Ricci tensor Ric , introduced in Chapter 7.G.

21. THEOREM (MYERS). Let M be an n -dimensional Riemannian manifold, and $\gamma: [0, L] \rightarrow M$ a geodesic parameterized by arclength. Let $r > 0$ be a constant, and suppose that

$$-\text{Ric}(\gamma'(t), \gamma'(t)) \geq \frac{n-1}{r^2} \quad \text{for all } t,$$

and that γ has length $L > \pi r$. Then there is a point $\tau \in (0, L)$ conjugate to 0, and γ is not of minimal length.

PROOF. Choose parallel, everywhere orthonormal vector fields Y_1, \dots, Y_n along γ with $Y_1 = V$. Let $W_i(t) = (\sin \pi t/L) Y_i(t)$. Then

$$\begin{aligned} E_{**}(W_i, W_i) &= - \int_0^L \left\langle W_i, \frac{D^2 W_i}{dt^2} + R(W_i, V)V \right\rangle dt \\ &= \int_0^L \left(\sin \frac{\pi t}{L} \right)^2 \left[\frac{\pi^2}{L^2} - \langle R(Y_i(t), Y_1(t)) Y_1(t), Y_i(t) \rangle \right] dt. \end{aligned}$$

Summing for $i = 2, \dots, n$, we obtain

$$\sum_{i=2}^n E_{**}(W_i, W_i) = \int_0^L \left(\sin \frac{\pi t}{L} \right)^2 \left[\frac{(n-1)\pi^2}{L^2} + \text{Ric}(Y_1(t), Y_1(t)) \right] dt.$$

By hypothesis, the term in brackets is < 0 , so some $E_{**}(W_i, W_i)$ is < 0 . Thus γ is not of minimal length, and there is $\tau \in (0, L)$ conjugate to 0 (same reasoning as in Corollary 20). ♦

Remark: If $\gamma: [0, L] \rightarrow S^n(r)$ is a geodesic parameterized by arclength, and Y is any parallel vector field along γ which is perpendicular to γ , then $W(t) = (\sin \pi t/L)Y(t)$ satisfies $E_{**}(W, W) < 0$. The vector fields W_i in the above proof come from such vector fields W by means of the map Φ of Lemma 18. This may make the proof of Myers' theorem somewhat less mysterious.

The next result reproduces the reasoning in the third part of Bonnet's theorem, together with an observation of interest only in the higher dimensional case.

22. COROLLARY. Let M be a complete connected n -dimensional manifold with

$$-\text{Ric}(X, X) \geq \frac{n-1}{r^2}$$

for all unit vectors X , where $r > 0$ is a constant. (This hypothesis holds, in particular, if $K(P) \geq 1/r^2$ for all plane sections P .) Then M is actually compact, with diameter $\leq \pi r$. Moreover, the fundamental group of M is finite.

PROOF. The proof of the first part is exactly the same as in Bonnet's theorem. To prove that the fundamental group of M is finite, we simply consider the universal covering space $\pi: \tilde{M} \rightarrow M$ of the Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$. Clearly $(\tilde{M}, \pi^*\langle \cdot, \cdot \rangle)$ is complete, and its Ricci curvature also satisfies $-\text{Ric}(X, X) \geq (n-1)/r^2$ for all unit vectors X . Therefore \tilde{M} is compact. ♦

Although Theorem 19 certainly generalizes Bonnet's theorem very nicely, we do lose some information in this approach. Roughly speaking, we have generalized to higher dimensions only the first part of the Sturm comparison theorem, telling us that our Jacobi field $\Phi(W)$ must vanish somewhere on (a, b) ; we have not generalized the second part by comparing $\|\Phi(W)\|$ with $\|W\|$ up to the first zero of $\Phi(W)$. Such information is provided by

23. THEOREM (THE RAUCH COMPARISON THEOREM). Let M_1 and M_2 be two manifolds of the same dimension n , and let $\gamma_i: [a, b] \rightarrow M_i$ be geodesics parameterized by arclength such that

- (1) no number $\tau \in (a, b]$ is a conjugate value of 0 along γ_1 in M_1 or along γ_2 in M_2 .

Let W_i be Jacobi fields along γ_i such that

- (2) $W_i(a) = 0$,

$$(3) \quad \left\| \frac{DW_1}{dt}(a) \right\| = \left\| \frac{DW_2}{dt}(a) \right\|,$$

(4) W_i is perpendicular to γ_i .

For all $t \in [a, b]$, suppose that for all 2-dimensional $P_i \subset (M_i)_{\gamma_i(t)}$, the curvatures K_i satisfy

$$(5) \quad K_1(P_1) \leq K_2(P_2).$$

Then

$$\|W_1(t)\| \geq \|W_2(t)\| \quad \text{for all } t \in [a, b].$$

PROOF. If $W_2 = 0$, the theorem is trivial. If W_2 is not the 0 vector field, then $W_2(t) \neq 0$ for all $t \in (a, b)$, since γ_2 has no conjugate points. Naturally, $W_1(t)$ is also non-zero for all $t \in (a, b)$. It obviously suffices to prove that

$$(1) \quad \lim_{t \rightarrow 0} \frac{\langle W_1, W_1 \rangle(t)}{\langle W_2, W_2 \rangle(t)} = 1$$

$$(2) \quad \frac{d}{dt} \frac{\langle W_1, W_1 \rangle(t)}{\langle W_2, W_2 \rangle(t)} \geq 0 \quad \text{for } t \in (a, b).$$

To prove (1) we note that

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{\langle W_1, W_1 \rangle(t)}{\langle W_2, W_2 \rangle(t)} &= \lim_{t \rightarrow 0} \frac{\left\langle W_1, \frac{DW_1}{dt} \right\rangle(t)}{\left\langle W_2, \frac{DW_2}{dt} \right\rangle(t)} && \text{by L'Hôpital's Rule} \\ &= \lim_{t \rightarrow 0} \frac{\left\langle \frac{DW_1}{dt}, \frac{DW_1}{dt} \right\rangle(t) + \left\langle W_1, \frac{D^2 W_1}{dt^2} \right\rangle(t)}{\left\langle \frac{DW_2}{dt}, \frac{DW_2}{dt} \right\rangle(t) + \left\langle W_2, \frac{D^2 W_2}{dt^2} \right\rangle(t)} \\ &&& \text{by L'Hôpital's Rule} \\ &= 1, && \text{by hypothesis (3).} \end{aligned}$$

(Note that the first use of L'Hôpital's Rule is a genuine one, necessitated by the fact that we need to look at $\langle W_i, W_i \rangle$, rather than $\|W_i\|$. The second use, however, represents the same trivial case which occurs in the Sturm comparison theorem.)

Equation (2) is equivalent to

$$\langle W_2, W_2 \rangle \cdot \left\langle W_1, \frac{DW_1}{dt} \right\rangle \geq \langle W_1, W_1 \rangle \cdot \left\langle W_2, \frac{DW_2}{dt} \right\rangle,$$

so for each $t_0 \in (a, b)$ it suffices to show that

$$(2') \quad \left\langle W_1, \frac{DW_1}{dt} \right\rangle(t_0) \geq c^2 \left\langle W_2, \frac{DW_2}{dt} \right\rangle(t_0),$$

where $c = \|W_1(t_0)\|/\|W_2(t_0)\|$.

But, since the W_i are Jacobi fields, and $W_i(a) = 0$, the second variation formula shows that

$$\left\langle W_i, \frac{DW_i}{dt} \right\rangle(t_0) = E_{**}(\tilde{W}_i, \tilde{W}_i),$$

where $\tilde{W}_i = W_i|_{[a, t_0]}$.

Therefore, we just have to prove that

$$(2'') \quad E_{**}(\tilde{W}_1, \tilde{W}_1) \geq c^2 E_{**}(\tilde{W}_2, \tilde{W}_2).$$

Consider the map Φ of Lemma 18, constructed for the geodesics $\gamma_i|_{[0, t_0]}$. Since $W_i(t_0)$ are both non-zero, and orthogonal to γ_i , we can obviously define Φ so that

$$(3) \quad \Phi(\tilde{W}_1)(t_0) = c \tilde{W}_2(t_0).$$

In the first part of the proof of Theorem 19 we showed that

$$(4) \quad E_{**}(\tilde{W}_1, \tilde{W}_1) \geq E_{**}(\Phi(\tilde{W}_1), \Phi(\tilde{W}_1)).$$

On the other hand, we have $\tilde{W}_2(a) = \Phi(\tilde{W}_1)(a) = 0$ by hypothesis (2) and the norm preserving property of Φ , while $c\tilde{W}_2(t_0) = \Phi(\tilde{W}_1)(t_0)$ by (3). So Corollary 10 yields

$$(5) \quad \begin{aligned} E_{**}(\Phi(\tilde{W}_1), \Phi(\tilde{W}_1)) &\geq E_{**}(c\tilde{W}_2, c\tilde{W}_2) \\ &= c^2 E_{**}(\tilde{W}_2, \tilde{W}_2). \end{aligned}$$

Equations (4) and (5) together give the required equation (2''). ♦

Unless you have become totally lost in these generalities, it should be clear that Theorem 23 can also be used to prove the results of Corollary 20. Rauch actually used his comparison theorem to prove a much more striking result, concerning “ δ -pinched” manifolds. These are Riemannian manifolds satisfying

$$\delta A \leq K(P) \leq A$$

for all 2-dimensional subspaces $P \subset M_p$ at all points $p \in M$; here A is a constant, which we can assume is 1 if we are willing to multiply the metric $\langle \cdot, \cdot \rangle$ by a constant. Rauch proved that if M is complete and simply-connected, and δ -pinched for a certain $\delta \sim .74$, then M is homeomorphic to a sphere. Improvements by Berger and Klingenberg have established the

SPHERE THEOREM. Let M be a complete, simply-connected Riemannian manifold of dimension n whose sectional curvatures $K(P)$ satisfy

$$\delta \leq K(P) \leq 1$$

for some constant $\delta > 1/4$. Then M is homeomorphic to S^n .

It is known that for even n this result breaks down if we allow $\delta = 1/4$. We will not go into the rather detailed proofs of this and related recent results, which together would make up a good sized monograph. A large selection is coherently presented in Gromoll, Klingenberg, Meyer [1], and references to a few more will be found in the bibliography.

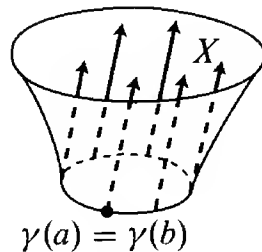
One of the most striking recent results completely clarifies the requirement in many of our theorems that the sectional curvatures should not only be positive, but also bounded away from 0. Naturally, this latter condition can fail only when the manifold M is not compact; but in this case the structure of M is completely determined:

THEOREM (GROMOLL-MEYER). If M is a connected, complete, non-compact n -dimensional manifold with all sectional curvatures positive, then M is diffeomorphic to \mathbb{R}^n .

Although we must omit the proofs of these theorems, we can prove an older and easier, but still very striking, result about the topology of manifolds whose sectional curvatures are all positive. We begin with a lemma that will also be used later on.

24. LEMMA (SYNGE). Let M be an orientable even-dimensional Riemannian manifold with all sectional curvatures positive. Let $\gamma: [a, b] \rightarrow M$ be a geodesic which is closed [that is, $\gamma(a) = \gamma(b)$ and $\gamma'(a) = \gamma'(b)$]. Then there is a variation $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ of γ such that all curves $\bar{\alpha}(u)$ are closed curves with length $\bar{\alpha}(u) < \text{length } \gamma$ for $u \neq 0$.

PROOF. Let $\mathcal{V} = \gamma'(0)^\perp \subset M_p$ be the $(n-1)$ -dimensional subspace of all $X_p \in M_p$ which are perpendicular to $\gamma'(0)$. Define $\phi: \mathcal{V} \rightarrow \mathcal{V}$ to be the result of parallel translation around γ . Then ϕ is a norm preserving linear transformation, with matrix A satisfying $AA^t = 1$ (where A^t is the transpose of A), so $\det \phi = \pm 1$. Moreover, since M is orientable, it is easy to see that $\phi: \mathcal{V} \rightarrow \mathcal{V}$ must be orientation preserving, so that $\det \phi = +1$. Since the dimension of \mathcal{V} is odd, the characteristic polynomial of ϕ has at least one real root, so ϕ has a real eigenvalue, which clearly must be ± 1 . Moreover, the complex eigenvalues occur in conjugate pairs $\lambda, \bar{\lambda}$ with $\lambda\bar{\lambda} > 0$. The number of real eigenvalues ± 1 is therefore odd, and their product is positive, so at least one must be $+1$. Consequently, ϕ leaves some vector field fixed: $\phi(X_p) = X_p$ for some $X_p \in \mathcal{V}$. This means that parallel translation of X_p around γ produces a vector field X along γ with $X(a) = X(b)$.



Let $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow M$ be a variation of γ with variation vector field $\partial\alpha/\partial u(0, t) = X(t)$. Since $X(a) = X(b)$, we can clearly choose α so that $\alpha(u, a) = \alpha(u, b)$ for all u , which means that each $\bar{\alpha}(u)$ is a closed curve. Applying the second variation formula, and remembering that $DX/dt = 0$, we find that

$$E_{**}(X, X) = - \int_a^b \langle X(t), R(X(t), V(t))V(t) \rangle dt$$

< 0 , by the hypothesis on sectional curvatures.

This means that for sufficiently small $u \neq 0$, the curves $\bar{\alpha}(u)$ have smaller energy than γ . ♦

With this Lemma we can easily prove the following result, provided that we accept an “intuitively obvious” fact, whose proof will come soon afterwards. We will temporarily use the term “closed path” for a continuous path $c: [0, 1] \rightarrow M$ with $c(0) = c(1)$; the term “smooth closed path” will be used for a smooth path $c: [0, 1] \rightarrow M$ with $c(0) = c(1)$ and $c'(0) = c'(1)$.

25. THEOREM (SYNGE). Let M be a compact, connected, orientable, even-dimensional Riemannian manifold with all sectional curvatures positive. Then M is simply-connected.

PROOF. Pick a point $p \in M$ and suppose that $\pi_1(M, p) \neq 0$. Let $c: [0, 1] \rightarrow M$ be a closed path with $c(0) = c(1) = p$, representing a non-zero element of $\pi_1(M, p)$. We say that a closed path γ is in the same **free homotopy class** as c if c and γ are homotopic, considered simply as maps from S^1 into M .

CLAIM. There is a closed curve $\gamma: [0, 1] \rightarrow M$ in the same free homotopy class as c which has smaller length than any other closed curve in this free homotopy class.

If we accept this claim, then it is clear that γ must be a smooth closed geodesic. For every sufficiently small segment of γ must coincide with a geodesic, since geodesics are the smallest paths between sufficiently close points.

The proof is now immediate, for we obtain a contradiction by applying Synge's Lemma to γ . ♦

Before we proceed with the proof of the Claim, we add a few remarks. The hypothesis that M is *compact* can be replaced by the hypothesis that M is complete and has sectional curvatures bounded away from 0, by Corollary 22. In fact, the Theorem of Gromoll-Meyer (page 239) shows that compactness can be replaced by completeness alone. The hypothesis that M is *orientable* is clearly necessary, as shown by the projective spaces P_n with n even. However, one can easily show (Problem 3) that if M is not orientable, then $\pi_1(M) \approx \mathbb{Z}_2$. The necessity of assuming that M is *even-dimensional* is shown by the projective spaces P_n with n odd. Without this assumption we must content ourselves with showing (Problem 3) that if M is a compact, connected odd-dimensional manifold with all sectional curvatures positive, then M is orientable.

We will now give two different proofs of the Claim. The first of these, the official proof, uses a few facts about covering spaces, and is generally considered to be quite elegant. The second proof is a more typical example of the sort of "direct methods" which one can sometimes use in order to establish that solutions to calculus of variation problems actually exist, instead of merely finding conditions on the presumed solution; it is similar to arguments first used by Hilbert for that sort of question (and similar arguments could be used to give an alternative demonstration that a minimal geodesic exists between any two points in a complete manifold). That, I feel, is one good reason for including it;

it also turns out that this proof is no harder than the first proof if the details are handled intelligently.

26. PROPOSITION. If $(M, \langle \cdot, \cdot \rangle)$ is a non-simply-connected compact Riemannian manifold, then every free homotopy class contains a curve of minimum length.

FIRST PROOF. Let \tilde{M} be the universal covering space of M and $\pi: \tilde{M} \rightarrow M$ the projection; the complete Riemannian metric $\pi^*\langle \cdot, \cdot \rangle$ on \tilde{M} gives an ordinary metric d on \tilde{M} . Recall that a homeomorphism $\phi: \tilde{M} \rightarrow \tilde{M}$ with $\pi \circ \phi = \pi$ is called a “covering transformation” or “deck transformation” of \tilde{M} . The set \mathcal{D} of all deck transformations is in one-one correspondence with $\pi_1(M, p)$ for any $p \in M$, and d is invariant under the action of \mathcal{D} .

Given a closed path $c: [0, 1] \rightarrow M$, let $\tilde{c}: [0, 1] \rightarrow \tilde{M}$ be a lifting, starting at some point $q \in \pi^{-1}(c(0))$. Then

$$\tilde{c}(1) = \delta(q) = \delta(\tilde{c}(0)) \quad \text{for a unique } \delta \in \mathcal{D}.$$

To see how this δ depends on the choice of $q \in \pi^{-1}(c(0))$, we note that any other point $q \in \pi^{-1}(c(0))$ is $\phi(q)$ for some $\phi \in \mathcal{D}$, and that the lifting \tilde{c} of c starting at $\phi(q)$ is just $\phi \circ \tilde{c}$. This means that

$$\tilde{c}(1) = \phi(\tilde{c}(1)) = \phi(\delta(q)) = \phi \circ \delta \circ \phi^{-1}(\phi(q)) = \phi \circ \delta \circ \phi^{-1}(\tilde{c}(0)).$$

Thus:

- (1) The conjugacy class $\{\phi\delta\phi^{-1} : \phi \in \mathcal{D}\}$ does not depend on the choice of $q \in \pi^{-1}(c(0))$.

We also claim:

- (2) If $c_1: [0, 1] \rightarrow M$ is freely homotopic to c , then it determines the same conjugacy class.

For, we have a map $H: I \times I \rightarrow M$ with

$$\begin{aligned} H(t, 0) &= c(t) \\ H(t, 1) &= c_1(t) \\ H(0, s) &= H(1, s). \end{aligned}$$

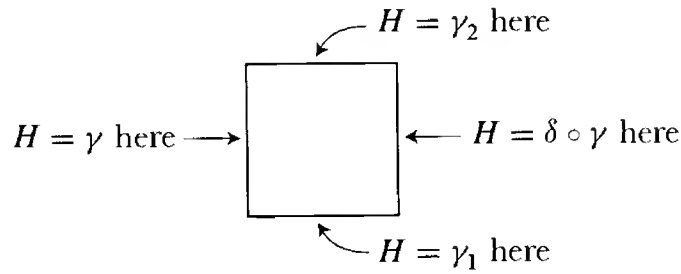
Let \tilde{H} be a lifting of H , and define $\tilde{c}(t) = \tilde{H}(t, 0)$ and $\tilde{c}_1(t) = \tilde{H}(t, 1)$. We have $\tilde{H}(1, s) = \delta(s)(\tilde{H}(0, s))$ for some $\delta(s)$, and all $\delta(s)$ must be the same δ , by

continuity. Thus both c and c_1 determine the same conjugacy class $\{\phi\delta\phi^{-1} : \phi \in \mathcal{D}\}$.

Finally, we claim:

- (3) If $\gamma_1, \gamma_2: [0, 1] \rightarrow \tilde{M}$ are paths with $\gamma_i(1) = \delta(\gamma_i(0))$, then $\pi \circ \gamma_1$ is freely homotopic to $\pi \circ \gamma_2$.

To prove this, we let $\gamma: [0, 1] \rightarrow \tilde{M}$ be a path from $\gamma_1(0)$ to $\gamma_2(0)$. Then $\delta \circ \gamma$ is a path from $\gamma_1(1)$ to $\gamma_2(1)$. So we can define a continuous map $H: \partial([0, 1] \times [0, 1]) \rightarrow \tilde{M}$ as follows:



Since \tilde{M} is simply-connected, we can extend this to a map $H: [0, 1] \times [0, 1] \rightarrow \tilde{M}$. Then $\pi \circ H: [0, 1] \times [0, 1] \rightarrow M$ satisfies

$$\pi \circ H(0, s) = \pi \circ H(1, s) \quad \text{for all } s \in [0, 1].$$

So $\pi \circ \gamma_1$ is freely homotopic to $\pi \circ \gamma_2$.

Now let $\{\phi\delta\phi^{-1} : \phi \in \mathcal{D}\}$ be the conjugacy class corresponding to our given free homotopy class, and define $h_\delta: \tilde{M} \rightarrow \tilde{M}$ by

$$h_\delta(q) = \inf\{d(q, \phi\delta\phi^{-1}(q)) : \phi \in \mathcal{D}\};$$

this is well-defined, since it clearly depends only on the conjugacy class of δ . Notice that for each q there is some ϕ (depending on q) such that

$$h_\delta(q) = d(q, \phi\delta\phi^{-1}(q));$$

this follows from the fact that \mathcal{D} acts discretely. It is also clear that h_δ is invariant under the action of \mathcal{D} . It follows that h_δ takes on its minimum on \tilde{M} ; for there is a compact set $K \subset \tilde{M}$ with $\pi(K) = M$, and consequently $\mathcal{D}(K) = \tilde{M}$, which means that the minimum of h_δ on K is also the minimum on all of \tilde{M} . Say that h_δ takes on its minimum at $q_0 \in \tilde{M}$, and that

$$h_\delta(q_0) = d(q_0, \phi_0\delta\phi_0^{-1}(q_0)).$$

Let γ be a minimal geodesic in \tilde{M} from q_0 to $\phi_0\delta\phi_0^{-1}(q_0)$, with length $\gamma = h_\delta(q_0)$. Then also

$$\text{length } \pi \circ \gamma = h_\delta(q_0).$$

The curve $\pi \circ \gamma$ is in the given free homotopy class, by (3). If c is any other curve in the free homotopy class, and \tilde{c} is any lifting, starting at some point q , then $\tilde{c}(1)$ is $\psi\delta\psi^{-1}(q)$ for some $\psi \in \mathcal{D}$, and consequently

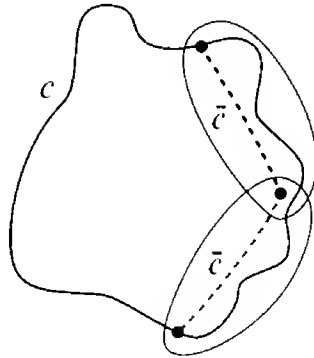
$$\begin{aligned} \text{length } c &= \text{length } \tilde{c} \geq d(q, \psi\delta\psi^{-1}(q)) \\ &\geq h_\delta(q) \geq h_\delta(q_0) = \text{length } \pi \circ \gamma. \end{aligned}$$

SECOND PROOF. Since M is compact, there is a finite open cover U_1, \dots, U_r of M by geodesically convex sets. By the Lebesgue covering lemma, there is $\varepsilon > 0$ such that any set A with diameter $< \varepsilon$ lies entirely in some U_α .

A closed curve c in M will be called **special** if there is a sequence $p_0, p_1, \dots, p_N = p_0$ of points in M such that

- (i) for each j , the points p_{j-1}, p_j both lie in some U_α
- (ii) c is the union of minimal geodesics c_j joining p_{j-1} to p_j .

Given an arbitrary closed curve $c: [0, 1] \rightarrow M$, there is always a special closed curve \bar{c} in the same free homotopy class as c , with length $l(\bar{c}) \leq l(c)$. To prove this, we consider the cover $\{c^{-1}(U_\alpha)\}$ of $[0, 1]$. The Lebesgue covering lemma implies the existence of a sequence $0 = t_0 \leq \dots \leq t_N = 1$ such that each $[t_{j-1}, t_j]$ is contained in some $c^{-1}(U_\alpha)$; this means that the restriction $c|_{[t_{j-1}, t_j]}$ is contained in U_α . We can then let \bar{c} be the union of the minimal geodesics



in U_α joining $p_{j-1} = c(t_{j-1})$ to $p_j = c(t_j)$. It is clear that $l(\bar{c}) \leq l(c)$. Since each U_α is geodesically convex, it is also clear that \bar{c} is homotopic to c .

Now consider a particular free homotopy class of closed curves. The set of lengths of all closed curves in this free homotopy class has a greatest lower bound $l \geq 0$. Our aim is to find a closed curve c in this free homotopy class

with length $l(c) = l$. We can certainly find a sequence $c^{(i)}$ of curves in this free homotopy class with

$$(1) \quad l(c^{(i)}) \rightarrow l;$$

we might as well assume that we also have

$$(2) \quad l(c^{(i)}) < 2l \quad \text{for all } i.$$

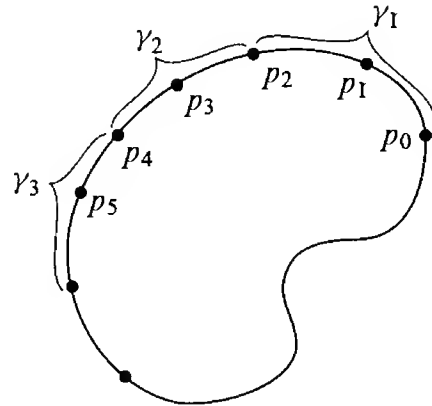
Finally, we can clearly assume that

$$(3) \quad \text{each } c^{(i)} \text{ is special.}$$

Now in the definition of a special closed curve, no bound was placed on the number N of division points involved. However, if the N for any of our curves $c^{(i)}$ is sufficiently large, then we can always find a new $c^{(i)}$, in the same homotopy class, and with no larger length, but with a smaller N . To see why this is so, consider the $[N/2]$ curves

$$\gamma_1 = c^{(i)} \text{ from } p_0 \text{ to } p_2$$

$$\gamma_2 = c^{(i)} \text{ from } p_2 \text{ to } p_4$$

$$\vdots$$


If any one of these curves has length $< \varepsilon$, then it lies entirely in some U_α , so we can replace it by a single minimal geodesic, thereby reducing N . Clearly:

if $l(c) < \left[\frac{N}{2}\right] \varepsilon$, then some γ_v has length $< \varepsilon$, so N can be reduced.

Using (2), we find:

if $2l < \left[\frac{N}{2}\right] \varepsilon$, then some γ_v has length $< \varepsilon$, so N can be reduced.

Phrasing this slightly differently, we have:

if $N > 2\left(\frac{2l}{\varepsilon} + 1\right)$, then N can be reduced.

Since this is true for all curves $c^{(i)}$, we can assume that all curves $c^{(i)}$ have $N \leq 2(2l/\varepsilon + 1)$. Since extra points can always be stuck in, we can actually assume that

$$(4) \quad \text{each } c^{(i)} \text{ is special with } N = N_0 = \left\lceil 2\left(\frac{2l}{\varepsilon} + 1\right) \right\rceil.$$

The remainder of the proof is now very simple. Let

$$p_0^{(i)}, p_1^{(i)}, \dots, p_{N_0}^{(i)} = p_0^{(i)}$$

be the points determining $c^{(i)}$. Since M is compact, we may assume, by taking subsequences, that for each $j = 0, \dots, N_0$ we have

$$\lim_{i \rightarrow \infty} p_j^{(i)} = p_j \in M.$$

Joining the pairs p_{j-1}, p_j by minimal geodesics, we obtain a closed curve c . Clearly

$$\begin{aligned} l(c) &= \sum_{j=1}^{N_0} d(p_{j-1}, p_j) = \lim_{i \rightarrow \infty} \sum_{j=1}^{N_0} d(p_{j-1}^{(i)}, p_j^{(i)}) \\ &= \lim_{i \rightarrow \infty} l(c^{(i)}) = l. \end{aligned}$$

To prove that c is in the same homotopy class as the $c^{(i)}$, we will carry out the construction a wee bit more carefully. We assume first that the original choice of the U_α was made so that there are geodesically convex sets $W_\alpha \supset \overline{U_\alpha}$. Now consider a fixed j . For each i , the points $p_{j-1}^{(i)}, p_j^{(i)}$ both lie in some U_α . Since there are only finitely many U_α , one of them, $U_{\alpha(j)}$ say, must contain both $p_{j-1}^{(i)}$ and $p_j^{(i)}$ for infinitely many i . By taking a subsequence, we can assume that all $p_{j-1}^{(i)}$ and $p_j^{(i)}$ are in $U_{\alpha(j)}$. There are only finitely many j to consider, so by taking subsequences we may assume that

$$\text{all } p_{j-1}^{(i)} \text{ and } p_j^{(i)} \text{ are in some } U_{\alpha(j)}, \quad j = 1, \dots, N_0.$$

This clearly implies that

$$p_{j-1} \text{ and } p_j \text{ are in } \overline{U_{\alpha(j)}} \subset W_{\alpha(j)} \quad j = 1, \dots, N_0.$$

Using geodesic convexity of the W_α , it is easy to see that c is homotopic to any $c^{(i)}$. ♦

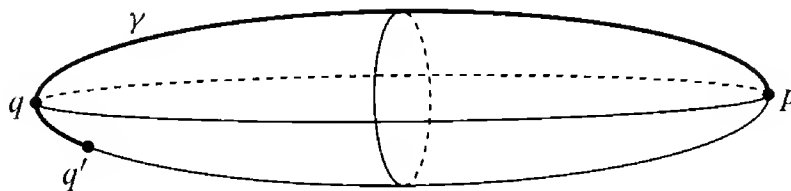
[We now reinstate the normal terminology, and use “closed curves” for curves $c: [0, 1] \rightarrow M$ with $c(0) = c(1)$ and $c'(0) = c'(1)$.]

We will end this chapter by considering a natural problem of deceptively simple appearance, which to this day remains unsolved. This problem will lead us to the study of “cut points”, which are related to, but still quite different from,

the conjugate points which we have been considering all along. We have seen (Corollary 22) that a complete connected manifold with all sectional curvatures $\geq 1/r^2$ has diameter $\leq \pi r$. It is natural to assume that in a similar way, a complete manifold with small sectional curvatures should have large diameter—if all $K(P) \leq 1/r^2$, then M should have diameter $\geq \pi r$. A counterexample to this conjecture is provided by projective space P_n , with constant curvature $= 1$, and diameter only $\pi/2$. And clearly, the larger the fundamental group, the smaller we might expect the diameter to be. An extreme case is represented by the torus, with infinite fundamental group. If we give the torus a flat metric, then $K \leq 1/r^2$ for every $r > 0$; on the other hand, we can also arrange for the diameter to be as small as we like. With the added hypothesis of simple connectivity, the conjecture still seems reasonable:

A complete, simply-connected, manifold with all $K(P) \leq 1/r^2$ should have diameter $\geq \pi r$.

One might expect to construct a proof of this conjecture along the following lines. We choose two points $p, q \in M$ at maximum distance apart, and consider a minimal geodesic $\gamma: [0, L] \rightarrow M$ which joins them. If γ has length $L < \pi r$, then we extend γ to $\bar{\gamma}: [0, L'] \rightarrow M$ with $L < L' < \pi r$, and the extended geodesic $\bar{\gamma}$ has no conjugate points, since $K(P) \leq 1/r^2$. Thus $\bar{\gamma}$ is a local minimum for length. Since p and q were already at the maximum distance apart, we might expect a contradiction to emerge from this construction. Of course, it can't, because we haven't used simple connectivity anywhere. The case of an ellipsoid shows where the problem lies. If γ is a geodesic joining the



two furthest points p and q , and extending somewhat further beyond q to q' , then γ is certainly not the shortest path between p and q' . But it *is* the shortest path among nearby paths, since γ contains no conjugate points. It seems clear that there is little hope of attacking this problem if we consider only conjugate points, since they only give us information about the *local* length minimizing property of geodesics, and our problem is a global one.

The notion of a cut point was made precisely in order to deal with global minimizing properties of geodesics. For simplicity, we will deal only with the case of a complete Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$, and we will let $d: M \times M \rightarrow \mathbb{R}$ be the ordinary metric on M determined by the Riemannian metric $\langle \cdot, \cdot \rangle$. Suppose we have a geodesic $\gamma: [0, \infty) \rightarrow M$ starting at a point $p = \gamma(0)$ in M , and parameterized by arclength. Consider the set

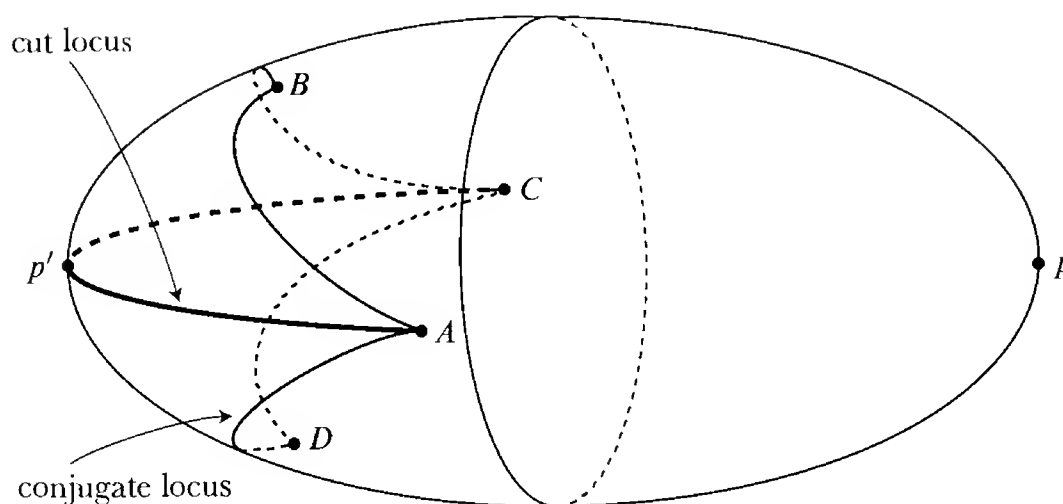
$$\begin{aligned} A &= \{t > 0 : d(p, \gamma(t)) = t\} \\ &= \{t > 0 : \gamma|_{[0, t]} \text{ is a minimal geodesic}\}. \end{aligned}$$

It is clear that either $A = (0, \infty)$ or else A is a set of the form $(0, a]$. If $A = (0, a]$, we say that $\gamma(a)$ is the **cut point** of p along the geodesic γ , while if $A = (0, \infty)$, we say that p has no cut point along γ . The **cut locus** $C(p) \subset M$ of p is then defined to be the set of all points which are cut points of p along some arclength parameterized geodesic starting from p . We also define the **cut locus** $\tilde{C}(p)$ of p in M_p to be the set of all vectors $aX \in M_p$ for which X is a unit vector and $\exp aX$ is the cut point of p along the geodesic $\gamma_X(t) = \exp tX$. Thus $C(p) = \exp(\tilde{C}(p))$. On the other hand, we define the **conjugate locus** of p in M_p to be the set of all vectors $aX \in M_p$ for which X is a unit vector and a is the first conjugate value of 0 along γ_X . A particular ray in M_p may contain neither a point of the conjugate locus nor a point of the cut locus. But if it contains a point aX of the conjugate locus, then by Theorem 6 it must also contain a point $a'X$ of the cut locus, with $a' \leq a$; briefly expressed, the cut point comes before or at the first conjugate point.

Notice that if M is compact, then there is certainly a cut point along every geodesic; but there may not be any conjugate points, as is shown by the case of a compact surface of everywhere negative curvature.

Suppose now that M is a simply-connected compact Riemannian manifold with all sectional curvatures $K(P) \leq 1/r^2$. If there is any point $p \in M$ for which the cut locus $\tilde{C}(p)$ in M_p and the conjugate locus in M_p intersect, say at a vector $v_p \in M_p$, then the diameter of M_p must be $\geq \pi r$. For, on the one hand, the geodesic $t \mapsto \exp tv_p$ is a minimal geodesic from p to $q = \exp v_p$, so $d(p, q) = \|v_p\|$; and on the other hand, the point q is conjugate to p , so $\|v_p\| \geq \pi r$ by Corollary 20.

Notice that the point q need not necessarily be the point furthest from p . For example, in the figure at the top of the next page, demonstrating the case of the ellipsoid on page 221, the cut locus of p is the portion of the geodesic from A to p' to C ; so q is A or C .

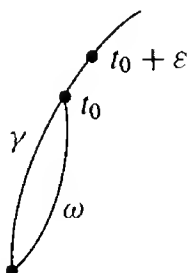


Unfortunately, it is not known whether such a point p always exists on a simply-connected compact manifold M . There are only partial results in this direction, and before giving one of them we will need to develop some basic properties of cut points.

One simple remark is sufficiently important to list as a separate result. Suppose that $\gamma: [0, \infty) \rightarrow M$ is a geodesic and $q = \gamma(t_0)$ comes *strictly before* the first cut point. Then, of course, any other geodesic ω from $p = \gamma(0)$ to q must have length $\omega \geq t_0$. But actually the strict inequality holds:

27. PROPOSITION. Let M be complete, let $\gamma: [0, \infty) \rightarrow M$ be a geodesic parameterized by arclength, and let $\gamma(t_0)$ come strictly before the cut point $\gamma(a)$ (if there is one). Then any other geodesic ω from $p = \gamma(0)$ to $q = \gamma(t_0)$ has length $\omega > t_0$.

PROOF. Suppose length $\omega = t_0 = \text{length } \gamma|[0, t_0]$. Choose $\varepsilon > 0$ so that $\gamma|[0, t_0 + \varepsilon]$ is also minimal. Then $\gamma|[0, t_0 + \varepsilon]$ has the same length as ω followed by $\gamma|[t_0, t_0 + \varepsilon]$. But this compound curve has a corner, so it can be made



shorter, and therefore $\gamma|[0, t_0 + \varepsilon]$ is *not* of minimal length, a contradiction. ♦

Notice that this argument does not work if $\gamma(t_0)$ is the cut point. In fact,

28. PROPOSITION. Let M be complete and let $\gamma: [0, \infty) \rightarrow M$ be a geodesic parameterized by arclength, with cut point $\gamma(a)$. Then at least one of the following holds:

- (1) The number a is the first conjugate value of 0 along γ .
- (2) There are at least two minimal geodesics from $p = \gamma(0)$ to $q = \gamma(a)$.

PROOF. Choose a sequence $a_1 > a_2 > a_3 > \dots$ with

$$(1) \quad \lim_{i \rightarrow \infty} a_i = a.$$

Let $b_i = d(p, \gamma(a_i)) < a_i$ and let X_i be unit vectors in M_p such that

$$t \mapsto \exp tX_i \quad 0 \leq t \leq b_i$$

is a minimal geodesic from p to $\gamma(a_i)$. Naturally, all X_i are distinct from $X = \gamma'(0)$. Then we also have

$$(2) \quad \lim_{i \rightarrow \infty} b_i = \lim_{i \rightarrow \infty} d(p, \gamma(a_i)) = d(p, \gamma(a)) = a.$$

Equation (2) shows that the vectors $b_i X_i$ are contained in a compact subset of M_p . Choosing a subsequence if necessary, we can assume that

$$(3) \quad \lim_{i \rightarrow \infty} b_i X_i = aY, \quad Y \in M_p \text{ a unit vector.}$$

Since $\exp aY = \lim_{i \rightarrow \infty} \exp b_i X_i = \lim_{i \rightarrow \infty} \gamma(a_i) = \gamma(a)$, the geodesic

$$t \mapsto \exp tY \quad 0 \leq t \leq a$$

is a minimal geodesic from p to q . So if $X \neq Y$ we have situation (2). To complete the proof we just have to show that if $X = Y$, then the number a must be conjugate to 0.

Now if $X = Y$, then $\lim_{i \rightarrow \infty} b_i X_i = aY = aX = \lim_{i \rightarrow \infty} a_i X$. But

$$(4) \quad \exp(b_i X_i) = \gamma(a_i) = \exp(a_i X).$$

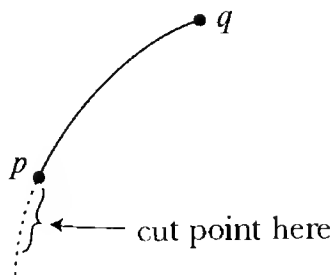
So every neighborhood of aX contains infinitely many pairs $b_i X_i, a_i X$ on which \exp has the same value; and these vectors $a_i X$ and $b_i X_i$ are definitely distinct

(since the X_i are different from X , or, just as conclusively, since $b_i < a < a_i$). So aX must be a critical point of \exp . Thus Theorem 7 shows that a is a conjugate value of 0 along γ . ♦

Proposition 28 can be used to derive several other facts about cut points. First of all, we have

29. PROPOSITION. In a complete manifold M , if q is the cut point of p along a geodesic γ from p to q , then p is the cut point of q along the geodesic $\bar{\gamma}$ obtained by traversing γ in the opposite direction.

PROOF. The hypothesis implies that γ is a minimal geodesic from p to q . So $\bar{\gamma}$ is minimal from q to p ; consequently, the cut point of $\bar{\gamma}$, if there is one, occurs *past* or at p . Now γ must satisfy one of the two alternatives in Proposition 28.



- (1) If q is conjugate to p along γ , then of course p is conjugate to q along $\bar{\gamma}$. The cut point must then occur *before* or at p . So it must occur at p .
- (2) If there is another minimizing geodesic from q to p , then again p must be the cut point, since Proposition 27 shows that there cannot be another minimal geodesic to a point strictly before the cut point. ♦

Consider now the “sphere bundle” $S(M)$ of M , consisting of all unit tangent vectors at all points of M ; this is a submanifold of the tangent bundle TM . Let $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ be the real numbers together with some other set “ ∞ ”. The ordering $<$ on \mathbb{R} can be extended to \mathbb{R}^* by defining $a < \infty$ for all $a \in \mathbb{R}$. We give \mathbb{R}^* the order topology (a basis consists of all sets of the form $(a, b) \subset \mathbb{R}$, together with all sets of the form $(a, \infty] = (a, \infty) \cup \{\infty\}$.) We now define a function $\mu: S(M) \rightarrow \mathbb{R}^*$ by

$$\mu(X) = \begin{cases} a > 0 & \text{if } aX \text{ is the cut point of } p \text{ along} \\ & \text{the geodesic } \gamma_X(t) = \exp tX \\ \infty & \text{if } \gamma_X \text{ has no cut point.} \end{cases}$$

30. THEOREM. If M is a complete manifold, then the function $\mu: S(M) \rightarrow \mathbb{R}^*$ is continuous.

PROOF. Let X_1, X_2, X_3, \dots be a sequence of unit vectors in $S(M)$ converging to a unit vector $X \in M_p$, and suppose that $a_i = \mu(X_i)$ did not converge to $a = \mu(X)$. Since the values of μ lie in the compact set $\{\alpha \in \mathbb{R}^* : \alpha \geq 0\}$, we can assume, by choosing a subsequence, that a_i converges to some $\alpha \in \mathbb{R}^*$ with $\alpha \neq a$. Suppose for the moment that α is in \mathbb{R} (and consequently all but finitely many a_i are in \mathbb{R}). Then $a_i X_i$ converges to αX . Now it is clear from the definition of μ that

$$d(p, \exp a_i X_i) = a_i.$$

So

$$\begin{aligned} d(p, \exp \alpha X) &= d(p, \lim_{i \rightarrow \infty} \exp a_i X_i) \\ &= \lim_{i \rightarrow \infty} d(p, \exp a_i X_i) \\ &= \lim_{i \rightarrow \infty} a_i = \alpha. \end{aligned}$$

This shows that the geodesic $t \mapsto \exp tX$ is minimizing on $[0, \alpha]$, and consequently $a = \mu(X) \geq \alpha$. If $\alpha = \infty$, it is easy to see that we must again have $a \geq \alpha$. So in order to derive a contradiction from the assumption that $a \neq \alpha$, we can assume that $a > \alpha$. Thus we are assuming that the vectors $a_i X_i$ approach the vector αX with $\alpha < a$. This means, in particular, that \exp_* is not singular at αX , since a conjugate point cannot come before a cut point.

By choosing a subsequence of our sequence, we can assume that either each $\gamma_i(t) = \exp(ta_i X_i)$ satisfies (1) of Proposition 28, or else that each γ_i satisfies (2). If each γ_i satisfies (1), then \exp_* is singular at each $a_i X_i$. Hence \exp_* is singular at $\alpha X = \lim_{i \rightarrow \infty} a_i X_i$, a contradiction.

If each γ_i satisfies (2), then there are unit vectors $Y_i \neq X_i$ such that $\exp(a_i Y_i) = \exp(a_i X_i)$. Since \exp is a diffeomorphism on some open neighborhood U of αX , these vectors $a_i Y_i$ must lie outside U . By choosing a subsequence, we can assume that Y_i approach a unit vector Y at p . Clearly Y also lies outside U , so $Y \neq X$. But

$$\begin{aligned} d(p, \exp \alpha Y) &= \lim_{i \rightarrow \infty} d(p, \exp a_i Y_i) \\ &= \lim_{i \rightarrow \infty} d(p, \exp a_i X_i) \\ &= \lim_{i \rightarrow \infty} a_i = \alpha. \end{aligned}$$

This shows that $t \mapsto \exp tY$ is another minimal geodesic from p to $\exp(\alpha X)$. Since $\exp(\alpha X)$ comes before the cut point, this cannot occur, according to Proposition 27. ♦

As a particular consequence of Theorem 30, the map $\mu: M_p \rightarrow \mathbb{R}^*$ is continuous for each $p \in M$. Therefore the set

$$E(p) = \{tv : v \in M_p \text{ is a unit vector and } 0 \leq t < \mu(v)\}$$

is clearly homeomorphic to an open n -dimensional cell.

31. THEOREM. Let M be complete. Then $\exp: M_p \rightarrow M$ maps $E(p)$ diffeomorphically onto an open subset of M , and M is the disjoint union of $\exp E(p)$ and $C(p)$.

PROOF. Clearly \exp_* is one-one on $E(p)$, since there are no vectors $w \in E(p)$ with $\exp w$ conjugate to p . To see that \exp is one-one on $E(p)$, consider $w_1, w_2 \in E(p)$, with $\|w_1\| \leq \|w_2\|$, say. If we had $\exp w_1 = \exp w_2 = q$, then the geodesic $\omega(t) = \exp tw_1$ would have length from p to q less than or equal to that of the geodesic $\gamma(t) = \exp tw_2$. This contradicts Proposition 27, since q comes before the cut point of γ .

We next claim that $\exp E(p)$ and $C(p)$ are disjoint. If not, then there is $w \in E(p)$ and $u \in \tilde{C}(p)$ with $\exp w = \exp u = q$. If $\|u\| \leq \|w\|$, we have the same contradiction as before. If $\|w\| < \|u\|$ we still have a contradiction, for then $t \mapsto \exp tw$ would be a geodesic from p to q shorter than the geodesic $t \mapsto \exp tu$, which is minimal since $u \in \tilde{C}(p)$.

Finally, let q be any point of M . Then there is an arclength parameterized minimal geodesic $\gamma(t) = \exp tv$ from $p = \gamma(0)$ to $q = \gamma(a)$. Clearly $a \leq \mu(v)$. So $av \in E(p)$ or $av \in \tilde{C}(p)$. ♦

32. COROLLARY. If M is complete, and $p \in M$, then M is compact if and only if every geodesic through p has a cut point. In particular, if every geodesic through M has a conjugate point, then M is compact.

PROOF. We already know that if M is compact, then every geodesic through p has a cut point. On the other hand, if every such geodesic has a cut point, then $\tilde{C}(p) \subset M_p$ is homeomorphic to S^{n-1} , and $E(p) \cup \tilde{C}(p)$ is a compact set. So $M = \exp(E(p) \cup \tilde{C}(p))$ is also compact. ♦

One reason that the cut locus $C(p)$ is so important is that most of the topological properties of M are concentrated in $C(p)$. For it is easy to see that there is a deformation retraction of $M - \{p\}$ into $C(p)$ —we just push points of $\exp E(p) - \{p\}$ along geodesics through p until they hit $C(p)$. Thus the homotopy groups $\pi_k(C(p))$ and singular homology $H_k(C(p))$ groups are isomorphic

to $\pi_k(M - \{p\})$ and $H_k(M - \{p\})$, respectively; and there are well-known relations between these groups and $\pi_k(M)$ and $H_k(M)$.

Another simple consequence of Theorem 30 is:

33. COROLLARY. If M is complete, then the distance $d(p, C(p))$ between p and its cut locus is a continuous function of p .

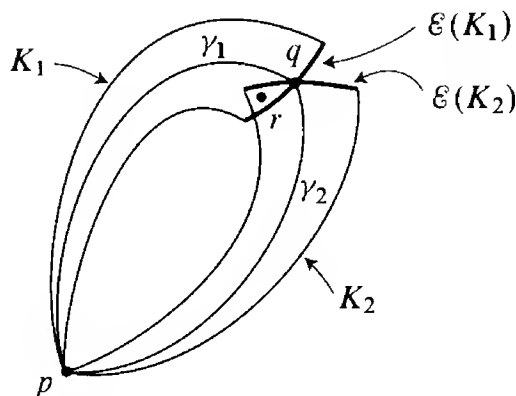
We are now beginning to approach our goal, although it may not look like it. We first prove the following important lemma, which improves on Proposition 28 when $\gamma(a)$ is a special point in $C(p)$.

34. LEMMA. Let p be a point in a complete manifold M and let q be a point of $C(p)$ closest to p . Then at least one of the following holds:

- (1) The point q is conjugate to p along some minimal geodesic from p to q .
- (2) There are exactly two minimal geodesics from p to q , and their tangent vectors at q are negatives of each other, so that together they give a geodesic beginning and ending at p .

PROOF. Suppose (1) does not hold. Then by Proposition 28 there are at least two minimal geodesics γ_1 and γ_2 from p to q . We will show that the tangent vectors of any two such γ_1 and γ_2 are negatives of each other at q ; this clearly implies in addition that there is not a third minimal geodesic γ_3 from p to q .

Let K_1 be a “cone” formed by the points on all geodesics of length $d(p, q)$ whose tangent vectors lie in a neighborhood of γ_1' at p ; and define K_2 similarly. The set $\mathcal{E}(K_1)$ of all the endpoints of the geodesics making up K_1 is a hypersurface containing q . If the tangent vectors of γ_1 and γ_2 are not negatives of each other at q , then $\mathcal{E}(K_1)$ crosses the corresponding hypersurface $\mathcal{E}(K_2)$.



It follows that there is a point r with

$$r \in [K_1 - \varepsilon(K_1)] \cap [K_2 - \varepsilon(K_2)].$$

Now r is joined to p by a geodesic $\bar{\gamma}_1$ lying in K_1 , and a geodesic $\bar{\gamma}_2$ lying in K_2 . Since q is the point of $C(p)$ closest to p , the point r must come strictly before the first conjugate point on both γ_1 and γ_2 . But this is impossible by Proposition 27. ♦

When the point p of Lemma 34 is very special we can say even more.

35. LEMMA. Let p be a point in a complete manifold M for which the distance $d(p, C(p))$ is smallest, and let q be a point of $C(p)$ closest to p . Suppose that q is not conjugate to p along a minimal geodesic from p to q . Then there is a closed geodesic made up of two minimal geodesics from p to q .

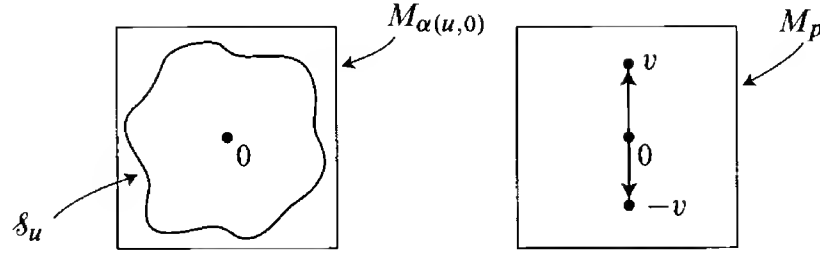
PROOF. Since q is a point of $C(p)$ closest to p , there are, by Lemma 34, exactly two minimal geodesics γ_1 and γ_2 from p to q , and their tangent vectors are negatives of each other at q . But our hypotheses imply also that p is a point of $C(q)$ closest to q . So there are also exactly two minimal geodesics from q to p , namely γ_1 and γ_2 again, and their tangent vectors are negatives of each other at p . ♦

36. THEOREM (KLINGENBERG). Let M be a compact simply-connected even-dimensional Riemannian manifold whose sectional curvatures satisfy $0 < K(P)$ for all 2-dimensional $P \subset M_q$, for all $q \in M$. Then for some point $p \in M$, the cut locus $\tilde{C}(p)$ in M_p and the conjugate locus in M_p intersect.

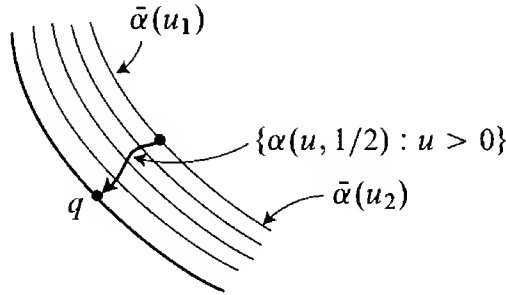
Consequently, if we also have $K(P) \leq 1/r^2$ for some $r > 0$, then M has diameter $\geq \pi r$.

PROOF. Let p be a point for which $d(p, C(p))$ has the smallest value, L say, and let $q \in C(p)$ be a point closest to p . We claim that q is conjugate to p along a minimal geodesic. Suppose it were not. Then by Lemma 35, there is a closed geodesic $\gamma: [0, 1] \rightarrow M$, of length $2L$, made up of two minimal geodesics from p to q . By Synge's Lemma, there is a variation $\alpha: [0, \varepsilon] \times [0, 1] \rightarrow M$ of α such that all $\bar{\alpha}(u)$ are closed curves of length $< 2L$ for $u > 0$. This means that for each $u > 0$, the set of points $\{\alpha(u, t)\}$ is the image under $\exp_{\alpha(u, 0)}$ of a

set \mathcal{S}_u in $E(\alpha(u, 0))$; this set \mathcal{S}_u is a closed curve, since $\exp_{\alpha(u, 0)}$ is a diffeomorphism on $E(\alpha(u, 0))$. But the points of γ are *not* all in $\exp_p E(p)$. Instead, the set $\{\gamma(t)\} - \{q\}$ is the image of a set in $E(p)$; this set consists of two open rays from $0 \in M_p$ to two vectors $v, -v \in M_p$. We will show that such a situation cannot arise.



Let $\mathcal{S} = \bigcup_{u>0} \mathcal{S}_u$, and consider the set C of all points in \mathcal{S} corresponding to points of the form $\alpha(u, 1/2)$ for $u > 0$. We claim that C is connected. This is



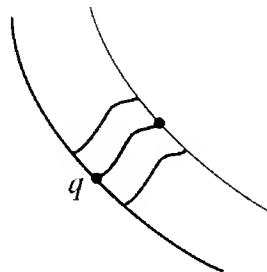
because the map

$$u \mapsto \exp_{\alpha(u, 0)}^{-1}(\alpha(u, 1/2))$$

is continuous, where $\exp_{\alpha(u, 0)}^{-1}$ denotes the inverse of the map $\exp_{\alpha(u, 0)} : E(\alpha(u, 0)) \rightarrow M$. Similarly, if C_n is the set of all points in \mathcal{S} corresponding to points of the form $\alpha(u, 1/2)$ for $0 < u < \frac{1}{n}$, then C_n is also connected.

Now consider the set B_n of points in \mathcal{S} corresponding to points of the form

$$\alpha(u, t) \quad \text{for } t \in \left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right) \quad \text{and } 0 < u < \frac{1}{n}.$$



The set B_n is also connected: it consists of the union of connected sets in δ_u for $0 < u < \frac{1}{n}$, and each of these contains a point of the connected set C_n .

Finally, consider the set

$$B = \bigcap_n \overline{B_n}.$$

As a decreasing intersection of compact, connected sets, it is also connected. It is clear that it is completely contained in M_p , and that it contains both v and $-v$. Therefore it must contain some other vector $w \in M_p$. But then it is easy to see that $t \mapsto \exp tw$ is another minimal geodesic from p to q , contradicting Lemma 34. ♦

PROBLEMS

1. Suppose that f satisfies the condition for second derivatives on page 204.
 - (a) Show that if f is composed with a suitable rotation, then the corresponding matrix $(\partial^2 f / \partial x_i \partial x_j)$ is diagonal (compare pg. II.50).
 - (b) Conclude that f has a local minimum at x .
 - (c) For $f(x, y) = (y - x^2)(y - 2x^2)$, show that f has a strict local minimum along every straight line through x , but that f does not have a local minimum at x .

2. (a) Prove the following “delicate Sturm comparison theorem”: Let f and h be two continuous functions $f \leq h$ on an open interval (a, b) , and let ϕ and η be two functions satisfying

$$\begin{aligned} (1) \quad & \phi'' + f\phi = 0 \\ (2) \quad & \eta'' + h\eta = 0 \end{aligned}$$

on (a, b) . Assume that $\phi(t) \neq 0$ for $t \in (a, b)$, and that

$$\lim_{t \rightarrow a^+} \phi(t) = \lim_{t \rightarrow b^-} \phi(t) = 0.$$

Then η must have a zero on (a, b) , unless $f = h$ everywhere on (a, b) and η is a constant multiple of ϕ on (a, b) .

(b) In the situation considered on page 231, let $V = d\gamma/dt$, and let Y be the unit vector field along $\gamma|(0, L)$ which is perpendicular to V and tangent to image α along $\gamma|(0, L)$. Let $W = fV + \phi Y$ be the decomposition of Proposition 3 for image α ; note that f and ϕ have (left- and right-hand) limits 0 at 0 and L . Conclude that $f = 0$ and that ϕ satisfies the hypotheses of part (a). Thus obtain a contradiction, demonstrating that we cannot have $L < \pi r$.

3. (a) Let M be a non-orientable C^∞ manifold. Let \tilde{M} be the set of all orientations μ_p for M_p , for all $p \in M$, and define $\pi: \tilde{M} \rightarrow M$ to be the map which takes each of the two orientations of M_p into p . Show that \tilde{M} has a natural C^∞ structure that makes $\pi: \tilde{M} \rightarrow M$ a 2-fold covering space of M , and that \tilde{M} is orientable.

(b) If M is a compact, connected, non-orientable, even-dimensional Riemannian manifold with all sectional curvatures positive, then $\pi_1(M) \approx \mathbb{Z}_2$.

(c) Let $c, \gamma: [0, 1] \rightarrow M$ be freely homotopic closed curves, and let $\tilde{c}, \tilde{\gamma}: [0, 1] \rightarrow \tilde{M}$ be curves with $\pi \circ \tilde{c} = c$ and $\pi \circ \tilde{\gamma} = \gamma$. Then $\tilde{c}(0) = \tilde{c}(1)$ if and only if $\tilde{\gamma}(0) = \tilde{\gamma}(1)$. Hence if M is non-orientable, then there is a closed curve $c: [0, 1] \rightarrow M$ of minimal length such that $\tilde{c}(0) \neq \tilde{c}(1)$.

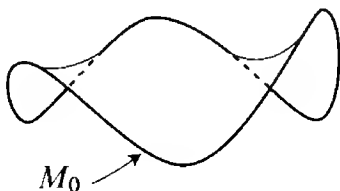
(d) If M is a compact, connected, odd-dimensional Riemannian manifold with all sectional curvatures positive, then M is orientable.

CHAPTER 9

VARIATIONS OF LENGTH, AREA, AND VOLUME

The classical calculus of variations was extended, quite soon after its inception, to deal with problems in several variables. In this chapter we will use these methods to study n -dimensional submanifolds $M \subset (N^m, \langle \cdot, \cdot \rangle)$ with minimal n -dimensional volume. Thus the material of this chapter may be regarded as a generalization of the study of geodesics which was carried out in Chapter I.9 and in Chapter 8 of this Volume. One big difference, aside from the greater difficulties to be encountered, is the fact that our results are truly extrinsic—all our theorems will be about the submanifolds of N , not about the structure of N itself.

When we look for curves which have the shortest length among all curves between 2 fixed endpoints, we find that the only possible candidates are geodesics (provided that we parameterize all curves proportionally to arclength). For the 2-dimensional analogue of this situation, we replace the two fixed endpoints in our Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$ by a compact 1-dimensional manifold M_0 (diffeomorphic to a finite union of circles). We then consider all immersed compact 2-dimensional manifolds-with-boundary M satisfying $\partial M = M_0$. Among these, we seek one which has minimum area; by the **area** of an immersed sur-



face $f: M \rightarrow N$ we mean the integral over M of the (2-dimensional) volume element dA determined by the induced metric $f^*\langle \cdot, \cdot \rangle$ (when M is oriented, we can consider dA to be a 2-form). Our approach to this problem will be similar to our approach in the analogous 1-dimensional case; we will find “critical points” for the area function. One important difference is that no particular parameterization of M will play a favored role.

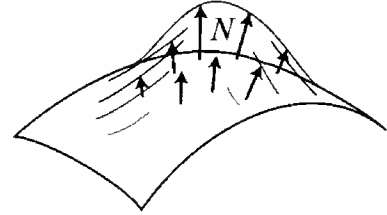
Before we try to find a general formula for the “variation of area”, we will first investigate the case of surfaces in \mathbb{R}^3 , which leads to an extraordinarily rich theory, of a very special sort. At first, we will not even consider general immersed surfaces-with-boundary, but only immersions $f: D \rightarrow \mathbb{R}^3$, where $D \subset \mathbb{R}^2$ is a compact 2-dimensional manifold-with-boundary. By a **variation** α of f we will mean a C^∞ function $\alpha: (-\varepsilon, \varepsilon) \times D \rightarrow \mathbb{R}^3$ with $\alpha(0, p) = f(p)$ for $p \in D$; for each $u \in (-\varepsilon, \varepsilon)$, we then define the function $\bar{\alpha}(u): D \rightarrow \mathbb{R}^3$ by $\bar{\alpha}(u)(p) = \alpha(u, p)$. Since $f = \bar{\alpha}(0)$ is an immersion, the same must be true of $\bar{\alpha}(u)$ for sufficiently small u (one needs compactness of D to prove this), so with no loss of generality we can assume that all $\bar{\alpha}(u)$ are immersions. As in the previous chapter, we define the **variation vector field** W by

$$W(p) = \frac{\partial \alpha}{\partial u}(0, p);$$

notice that $W(p) \in \mathbb{R}^3_{f(p)}$, so that W is a “vector field along f ”.

In almost every differential geometry book under the sun, the only variations considered are those of the form

$$(I) \quad \alpha(u, t_1, t_2) = f(t_1, t_2) + u \cdot \phi(t_1, t_2) \cdot N(t_1, t_2),$$



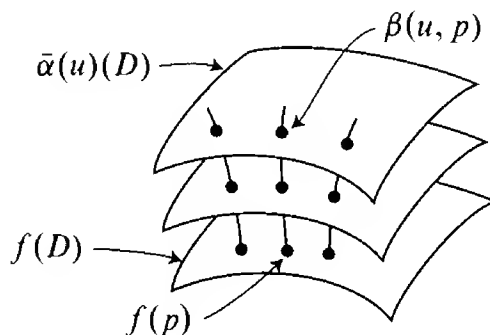
where $N(t_1, t_2)$ is the unit normal at $f(t_1, t_2)$, and ϕ is some C^∞ function. Thus α is a very special sort of “normal variation”—each curve $u \mapsto \alpha(u, t_1, t_2)$ is a straight line normal to the surface f , and the variation vector field W is just

$$W(t_1, t_2) = \phi(t_1, t_2) \cdot N(t_1, t_2).$$

The decision to ignore more general variations is partially justified by the following observations. In the first place, if we are given a variation $\alpha: (-\varepsilon, \varepsilon) \times D \rightarrow \mathbb{R}^3$ of f , then we can usually find a new variation $\beta: (-\varepsilon', \varepsilon') \times D \rightarrow \mathbb{R}^3$ of f such that

- (a) $\frac{\partial \beta}{\partial u}(0, p)$ is perpendicular to $f(D)$,
- (b) the surfaces $\bar{\alpha}(u)(D)$ and $\bar{\beta}(u)(D)$ are always the same, even though the parameterizations $\bar{\alpha}(u)$ and $\bar{\beta}(u)$ may be different.

To do this, we assume that the surfaces $\bar{\alpha}(u)(D)$ are all disjoint, and we consider the curves, parameterized by arclength, which are orthogonal to all the surfaces $\bar{\alpha}(u)(D)$. Then we let $\beta(u, p)$ be the unique point of $\bar{\alpha}(u)(D)$ which lies on



the curve passing through $f(p)$. Thus we can usually assume that our variation vector field W is perpendicular to $f(D)$. In the second place, when we take the derivative at 0 of the areas of the surfaces $\bar{\alpha}(u)(D)$, we naturally expect that the answer will depend only on W , exactly as in the case of arclength. If this expectation is correct, then we can even assume that α is of the form (1). This line of argument, intuitively reasonable as it may be, is perhaps not very satisfying. But we will have adequate opportunity to consider more general variations later on, when we re-examine surfaces immersed in an arbitrary Riemannian manifold. So for the time being, let us indulge in the classical simplification, which makes the calculations so much more manageable.

For the special variation given by (1) we have

$$(2) \quad \frac{\partial \alpha}{\partial t_i}(u, t_1, t_2) = \frac{\partial f}{\partial t_i} + u \cdot \frac{\partial \phi}{\partial t_i} \cdot N + u \cdot \phi \cdot \frac{\partial N}{\partial t_i} \quad [\text{all partials on the right evaluated at } (t_1, t_2)].$$

Let

$$(3) \quad g_{ij}(u)(t_1, t_2) = \left\langle \frac{\partial \alpha}{\partial t_i}(u, t_1, t_2), \frac{\partial \alpha}{\partial t_j}(u, t_1, t_2) \right\rangle,$$

so that the functions $g_{ij}(u)$ are the components of $\bar{\alpha}(u)^*(\ , \)$; in particular, then, $g_{ij} = g_{ij}(0)$ are the components of $f^*(\ , \)$. Since

$$\left\langle \frac{\partial f}{\partial t_i}, \frac{\partial N}{\partial t_j} \right\rangle = -l_{ij}, \quad \left\langle \frac{\partial f}{\partial t_i}, N \right\rangle = 0,$$

equations (2) and (3) give

$$g_{ij}(u) = g_{ij} - 2u\phi l_{ij} + u^2 a_{ij}(u),$$

where $(u, t_1, t_2) \mapsto a_{ij}(u)(t_1, t_2)$ is some continuous function. From this we obtain

$$\begin{aligned}\det g_{ij}(u) &= \det g_{ij} - 2u\phi[g_{11}l_{22} + g_{22}l_{11} - 2g_{12}l_{12}] + u^2b(u) \\ &= (\det g_{ij})[1 - 4u\phi H] + u^2b(u), \quad \text{by formula (B) of Chapter 3;}\end{aligned}$$

in this equation, $b(u)$ is a function having the same property as the $a_{ij}(u)$. It is now easy to see that

$$\left. \frac{\partial}{\partial u} \right|_{u=0} \det g_{ij}(u) = -4\phi H \det g_{ij},$$

from which we obtain

$$\left. \frac{\partial}{\partial u} \right|_{u=0} \sqrt{\det g_{ij}(u)} = -2\phi H \sqrt{\det g_{ij}}.$$

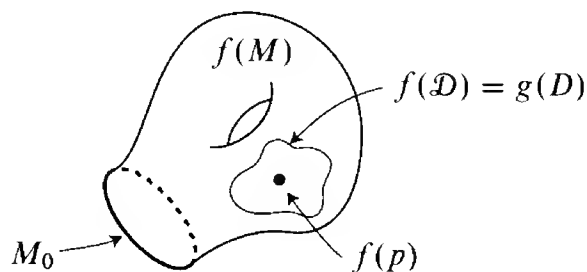
Let us denote by $A(\bar{\alpha}(u))$ the area of the immersed surface $\bar{\alpha}(u): D \rightarrow \mathbb{R}^3$. Then

$$\begin{aligned} (*) \quad \left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} &= \left. \frac{d}{du} \right|_{u=0} \int_D \sqrt{\det g_{ij}(u)} \, dt_1 \, dt_2 \\ &= \int_D \left. \frac{\partial}{\partial u} \right|_{u=0} \sqrt{\det g_{ij}(u)} \, dt_1 \, dt_2 \\ &= - \int_D 2\phi H \sqrt{\det g_{ij}} \, dt_1 \, dt_2 \\ &= - \int_D 2\phi H \, dA, \quad dA = \begin{array}{l} \text{volume element on } D \text{ for} \\ \text{the metric } f^*\langle \cdot, \cdot \rangle. \end{array}\end{aligned}$$

We are now ready to draw a conclusion.

1. PROPOSITION. Let M be a compact 2-dimensional manifold-with-boundary; and $f: M \rightarrow \mathbb{R}^3$ an immersion such that $f(\partial M)$ is a given compact 1-manifold $M_0 \subset \mathbb{R}^3$. If M is a critical point for the area function, among all such immersions, then M must be a minimal surface ($H = 0$ everywhere). In particular, if M has the minimum area among all such surfaces, then M is a minimal surface.

PROOF. Suppose that $H(p) \neq 0$ for some $p \in M$, say $H(p) > 0$. Choose a neighborhood \mathcal{D} of p so small that $H(q) > 0$ for all $q \in \mathcal{D}$. We can assume that $f(\mathcal{D})$ is also the image $g(D)$ for some immersion $g: D \rightarrow \mathbb{R}^3$ of a compact 2-dimensional manifold-with-boundary $D \subset \mathbb{R}^2$. Let $\phi: D \rightarrow \mathbb{R}$ be a C^∞



function which is ≥ 0 on D and $= 0$ in a neighborhood of ∂D . We can then define a variation α of f by letting

$$\begin{aligned} \alpha(u, p) &= f(p) & p \notin \mathcal{D} \\ \alpha(u, p) &= f(p) + u \cdot \phi(\bar{p}) \cdot N(\bar{p}), & \text{for } \bar{p} = g^{-1}(f(p)), \quad p \in \mathcal{D}. \end{aligned}$$

Formula (*) shows that

$$\left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_D 2\phi \tilde{H} dA,$$

where $\tilde{H}(t_1, t_2)$ is the mean curvature H at $f^{-1}(g(t_1, t_2))$. Since $\tilde{H} > 0$ everywhere on D , and since ϕ is ≥ 0 on D , but is not identically 0, the integral is positive; this is a contradiction. ♦

In the statement of Proposition 1 we have deliberately *not* claimed that a minimal surface actually is a critical point for the area function. We found that $H = 0$ is a *necessary* condition for a critical point by considering variations α which, first of all, vanish outside a small region, and, second of all, are normal to the surface. It is conceivable (well, just barely) that if we considered arbitrary variations, we would obtain another condition more stringent than $H = 0$. So we will have to wait a bit before we can assert with assurance that minimal surfaces are precisely the critical points for the area function. On the other hand, the second part of Proposition 1 is already the best we can hope for: among those surfaces with boundary M_0 , the one with minimum area must be a minimal surface; but we would not expect every minimal surface to have this property, any more than we expect every geodesic to be the shortest length between its endpoints.

The next result only begins to suggest how special minimal surfaces are.

2. PROPOSITION. Let M be an immersed surface in \mathbb{R}^3 with normal map $N: M \rightarrow S^2$. If M is minimal, then N is conformal (angle preserving) at all points where $K \neq 0$. Conversely, if N is conformal, and M is connected, then either M is a minimal surface, with $K < 0$ everywhere, or M is part of a sphere.

PROOF. Recall (Lemma II.7-20) that the map N is conformal at p if and only if there is $\mu(p) \neq 0$ such that

$$(1) \quad \langle N_* X_p, N_* Y_p \rangle = \mu(p) \langle X_p, Y_p \rangle \quad X_p, Y_p \in M_p.$$

We will make use of the third fundamental form III of M , which was defined in Chapter 2:

$$\begin{aligned} \text{III}(p)(X_p, Y_p) &= \langle N_* X_p, N_* Y_p \rangle \\ &= \langle N_*^2(X_p), Y_p \rangle, \quad \text{for } X_p, Y_p \in M_p. \end{aligned}$$

By Proposition 2-6 we have

$$(2) \quad \text{III} - 2H \cdot \text{II} + K \cdot \text{I} = 0.$$

Suppose first that M is minimal. Then (2) gives $\text{III} = -K \cdot \text{I}$, which shows that (1) holds with $\mu(p) = -K(p)$; hence N is conformal when $K(p) \neq 0$.

Conversely, suppose that N is conformal, so that it satisfies (1) for some function μ which is non-zero, and hence obviously positive. Then (2) gives

$$(K + \mu) \cdot \text{I} - 2H \cdot \text{II} = 0.$$

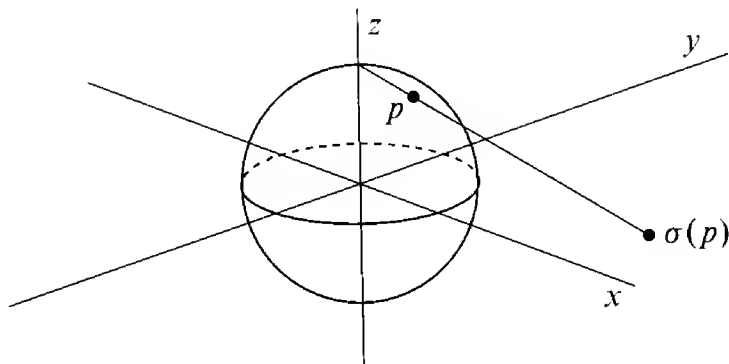
At a point p with $H(p) \neq 0$ we can therefore write $\text{II}(p)$ as a multiple of $\text{I}(p)$, which means that p is an umbilic. At a point p with $H(p) = 0$, we have $K(p) = -\mu(p) < 0$, so p cannot be an umbilic. In short,

$$p \text{ is an umbilic if and only if } H(p) \neq 0.$$

The set of umbilics is thus open. But it is also closed. So either: no points are umbilics, and $H = 0$ everywhere; or all points p are umbilics, and these umbilics are not flat points (since $H(p) \neq 0$), so M is part of a sphere. ♦

Back in Volume II, pg. 297, we mentioned that every 2-dimensional Riemannian manifold M is locally conformally equivalent to the plane: around each point p we can choose an “isothermal” coordinate system for which we have $g_{ij} = \mu \delta_{ij}$. Addendum I contains a proof of this result for general Riemannian 2-manifolds. On the other hand, Proposition 2 provides an easy way

of introducing isothermal coordinates around any non-flat point p of a minimal surface M . We need only find a conformal map $\sigma: S^2 - \{\text{point}\} \rightarrow \mathbb{R}^2$, and then $\sigma \circ N$ will be the required isothermal coordinate system in a neighborhood of p . But we already know such a conformal map σ , namely stereographic projection. It will be convenient to use the second version of stereographic projection, given on page 107. Recall that



$$\sigma(a, b, c) = \left(\frac{a}{1-c}, \frac{b}{1-c} \right)$$

$$\sigma^{-1}(x, y) = \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1} \right).$$

Naturally, we can find a conformal map $S^2 - \{q\} \rightarrow \mathbb{R}^2$ for any other point q merely by first rotating S^2 so that q goes to $(0, 0, 1)$.

Unfortunately, this method does not work at a flat point. To include such points we can, of course, appeal to the result of Addendum 1, valid for all surfaces. However, for minimal surfaces there is a considerably easier argument that still works at all points.

3. PROPOSITION. Isothermal coordinates can be introduced around any point of a minimal surface $M \subset \mathbb{R}^3$.

PROOF. We can assume that M is the graph of a function $h: U \rightarrow \mathbb{R}$, for $U \subset \mathbb{R}^2$, so that M is the image of the map $f(x, y) = (x, y, h(x, y))$. Introducing the classical notation

$$p = \frac{\partial h}{\partial x}, \quad q = \frac{\partial h}{\partial y},$$

$$r = \frac{\partial^2 h}{\partial x^2}, \quad s = \frac{\partial^2 h}{\partial x \partial y}, \quad t = \frac{\partial^2 h}{\partial y^2},$$

and using equation (B') on pg. III.137, we have

$$(1) \quad (1 + q^2)r - 2pqs + (1 + p^2)t = 0.$$

Setting

$$W = \sqrt{1 + p^2 + q^2},$$

we note that

$$\begin{aligned} \frac{\partial}{\partial x} \left(\frac{1 + q^2}{W} \right) - \frac{\partial}{\partial y} \left(\frac{pq}{W} \right) &= -\frac{p}{W^3} [(1 + q^2)r - 2pqs + (1 + p^2)t] \\ &= 0 \quad \text{by (1),} \end{aligned}$$

and similarly

$$\frac{\partial}{\partial x} \left(\frac{pq}{W} \right) - \frac{\partial}{\partial y} \left(\frac{1 + p^2}{W} \right) = 0.$$

This means that we can locally find functions α and β with

$$(2) \quad \begin{array}{ll} \text{(a)} \quad \frac{\partial \alpha}{\partial x} = \frac{1 + p^2}{W} & \text{(c)} \quad \frac{\partial \beta}{\partial x} = \frac{pq}{W} \\ \text{(b)} \quad \frac{\partial \alpha}{\partial y} = \frac{pq}{W} & \text{(d)} \quad \frac{\partial \beta}{\partial y} = \frac{1 + q^2}{W}. \end{array}$$

Consider the *transformation of Lewy*:

$$T(x, y) = (x + \alpha(x, y), y + \beta(x, y)).$$

Its Jacobian is

$$J(T)(x, y) = \begin{pmatrix} 1 + \frac{1 + p^2}{W} & \frac{pq}{W} \\ \frac{pq}{W} & 1 + \frac{1 + q^2}{W} \end{pmatrix},$$

with determinant

$$2 + \frac{2 + p^2 + q^2}{W} \geq 2.$$

So T has an inverse locally, and

$$\begin{aligned} J(T^{-1})(T(x, y)) &= [J(T)(x, y)]^{-1} \\ &= \frac{1}{\det J(T)(x, y)} \begin{pmatrix} 1 + \frac{1 + q^2}{W} & -\frac{pq}{W} \\ -\frac{pq}{W} & 1 + \frac{1 + p^2}{W} \end{pmatrix} \\ &= C \begin{pmatrix} 1 + W + q^2 & -pq \\ -pq & 1 + W + p^2 \end{pmatrix} \quad \text{for some } C. \end{aligned}$$

So

$$\begin{aligned}
 J(f \circ T^{-1})(T(x, y)) &= J(f)(x, y) \cdot J(T^{-1})(T(x, y)) \\
 &= C \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ p & q \end{pmatrix} \begin{pmatrix} 1 + W + q^2 & -pq \\ -pq & 1 + W + p^2 \end{pmatrix} \\
 &= C \cdot \begin{pmatrix} 1 + W + q^2 & -pq \\ -pq & 1 + W + p^2 \\ p + pW & q + qW \end{pmatrix}.
 \end{aligned}$$

It is easy to check that the two column vectors in this matrix are orthogonal, and that they have the same squared length

$$(1 + p^2 + q^2)(2W + 2 + p^2 + q^2).$$

Thus $f \circ T^{-1}$ is conformal, and its inverse is the desired isothermal coordinate system. ♦

The reader has probably noticed the similarity between this proof and the proof of Jörgens' Theorem (7-45). As a matter of fact, that proof of Jörgens' Theorem was motivated by manipulations with the minimal surface equation, and the original application of Jörgens' Theorem itself had been to reprove a result about minimal surfaces:

4. THEOREM (BERNSTEIN). Planes are the only minimal surfaces in \mathbb{R}^3 which are the graph of a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$.

PROOF. Suppose we have a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfying equation (1) in the previous proof. Then the functions α and β of equation (2) are defined on all of \mathbb{R}^2 (since \mathbb{R}^2 is simply-connected). From (b) and (c) of equation (2) we see that there is a function $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$\phi_x = \alpha \quad \text{and} \quad \phi_y = \beta.$$

Together with (a) and (d), we then have

$$\phi_{xx} = \frac{1 + p^2}{W}, \quad \phi_{xy} = \frac{pq}{W}, \quad \phi_{yy} = \frac{1 + q^2}{W},$$

which implies that

$$\phi_{xx}\phi_{yy} - (\phi_{xy})^2 = 1.$$

Jörgens' Theorem implies that

$$\frac{1+p^2}{W}, \quad \frac{pq}{W}, \quad \frac{1+q^2}{W}$$

are constants. A simple exercise then shows that p and q must be constants. ♦

The manipulations of the past few pages were undoubtedly unpleasant (not to say, slightly unmotivated), but they were really worth the trouble, because isothermal coordinates play such a vital role in the study of minimal surfaces.

5. PROPOSITION. If $f: M \rightarrow \mathbb{R}^3$ is a minimal immersion, and (u^1, u^2) is an isothermal coordinate system on M , then

$$\frac{\partial^2 f^i}{\partial u^1 \partial u^1} + \frac{\partial^2 f^i}{\partial u^2 \partial u^2} = 0 \quad i = 1, 2, 3.$$

Conversely, if this equation holds for a collection of isothermal coordinate systems covering M , then f is a minimal immersion.

PROOF. By equation (7) on page 136 we have

$$\Delta f = 2HN,$$

where N is the normal map, and Δ is the Laplacian. Therefore f is minimal if and only if $\Delta f^i = 0$ for $i = 1, 2, 3$. Since our coordinate system (u^1, u^2) is isothermal, Problem 7-23 shows that

$$\Delta f^i = \frac{1}{E} \left(\frac{\partial^2 f^i}{\partial u^1 \partial u^1} + \frac{\partial^2 f^i}{\partial u^2 \partial u^2} \right). \quad \diamond$$

Let us rephrase Proposition 5 just slightly. If $u = (u^1, u^2): U \rightarrow V \subset \mathbb{R}^2$ is an isothermal coordinate system on $U \subset M$, and $f: M \rightarrow \mathbb{R}^3$ is a minimal immersion, then each real-valued function $g^i = f^i \circ u^{-1}: V \rightarrow \mathbb{R}$ satisfies "Laplace's equation"

$$\frac{\partial^2 g^i}{\partial x^2} + \frac{\partial^2 g^i}{\partial y^2} = 0,$$

where $\partial/\partial x$ and $\partial/\partial y$ denote the ordinary partial derivatives in \mathbb{R}^2 . Now at this point complex analysis comes rushing in, waving its hands excitedly in its eagerness to enlighten us. It is a well-known result that locally any such function is the real part of a complex analytic function; we recall the argument briefly:

Suppose that g satisfies Laplace's equation

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = 0,$$

which we can also write as

$$\frac{\partial\left(\frac{\partial g}{\partial x}\right)}{\partial x} = \frac{\partial\left(-\frac{\partial g}{\partial y}\right)}{\partial y}.$$

According to Proposition I.6-0, there is locally a function h such that

$$\frac{\partial h}{\partial x} = -\frac{\partial g}{\partial y} \quad \frac{\partial h}{\partial y} = \frac{\partial g}{\partial x}.$$

But these are just the Cauchy-Riemann equations for $g + ih$, showing that this function is complex analytic, with real part $\operatorname{Re}(g + ih) = g$. The converse is even easier: If g is the real part of a complex analytic function $g + ih$, then the Cauchy-Riemann equations immediately lead to Laplace's equation for g .

A minimal surface M can thus be represented locally by

$$(x, y) \mapsto \Phi(x, y) = (\operatorname{Re} \phi_1(x + iy), \operatorname{Re} \phi_2(x + iy), \operatorname{Re} \phi_3(x + iy)) \in \mathbb{R}^3,$$

where the ϕ_i are complex analytic functions, and Φ itself is the inverse of an isothermal coordinate system. As one consequence of this representation, we see that every minimal surface in \mathbb{R}^3 is *automatically* real analytic (C^ω).

The fact that Φ^{-1} is an isothermal coordinate system can just as well be expressed by saying that Φ is conformal, and hence by the following two equations for the vectors $\partial\Phi/\partial x, \partial\Phi/\partial y \in \mathbb{R}^3$:

$$\left\langle \frac{\partial\Phi}{\partial x}, \frac{\partial\Phi}{\partial x} \right\rangle = \left\langle \frac{\partial\Phi}{\partial y}, \frac{\partial\Phi}{\partial y} \right\rangle \quad \left\langle \frac{\partial\Phi}{\partial x}, \frac{\partial\Phi}{\partial y} \right\rangle = 0.$$

Since the complex derivative ϕ_k' is given by

$$\begin{aligned} \phi_k'(x + iy) &= \frac{\partial \operatorname{Re} \phi_k}{\partial x} + i \frac{\partial \operatorname{Im} \phi_k}{\partial x} \\ &= \frac{\partial \operatorname{Re} \phi_k}{\partial x} - i \frac{\partial \operatorname{Re} \phi_k}{\partial y} \\ &= \frac{\partial \Phi^k}{\partial x} - i \frac{\partial \Phi^k}{\partial y}, \end{aligned}$$

our pair of equations for Φ is equivalent to the one complex equation $\sum_k (\phi_k')^2 = 0$; in terms of the functions $\psi_k = \phi_k'$ we can thus write our conditions as

$$\psi_1^2 + \psi_2^2 + \psi_3^2 = 0.$$

Now we can describe the solutions of this equation explicitly.

6. LEMMA. Let $V \subset \mathbb{C}$ be open, let g be meromorphic in V , and let f be analytic in V with a zero of order at least $2m$ at each point where g has a pole of order m . Then the functions

$$\psi_1 = \frac{1}{2}f(1 - g^2), \quad \psi_2 = \frac{i}{2}f(1 + g^2), \quad \psi_3 = fg$$

are analytic in V and satisfy $\psi_1^2 + \psi_2^2 + \psi_3^2 = 0$. Conversely, every triple ψ_1, ψ_2, ψ_3 of analytic functions satisfying $\psi_1^2 + \psi_2^2 + \psi_3^2 = 0$ on V can be represented this way.

PROOF. The first half of the Lemma is a direct calculation. Suppose, conversely, that we are given functions ψ_i satisfying the equation $\psi_1^2 + \psi_2^2 + \psi_3^2 = 0$, which we can also write in the form

$$(1) \quad (\psi_1 - i\psi_2)(\psi_1 + i\psi_2) = -\psi_3^2.$$

If ψ_3 is the 0 function, we choose $g = 0$ and $f = 2\psi_1$. If ψ_3 is not the 0 function, then $\psi_1 - i\psi_2$ is also not the 0 function, so we can define

$$(2) \quad f = \psi_1 - i\psi_2, \quad g = \frac{\psi_3}{\psi_1 - i\psi_2},$$

with f analytic and g meromorphic. Then equation (1) gives

$$(3) \quad \psi_1 + i\psi_2 = \frac{-\psi_3^2}{\psi_1 - i\psi_2} = -fg^2.$$

Equation (3) together with the definition of f in equation (2) shows that the ψ_i have the desired form. Equation (3) also shows that fg^2 is analytic, so f must have a zero of order at least $2m$ at each point where g has a pole of order m . ♦

It is now a simple matter to give a representation of minimal surfaces, due to Enneper and Weierstrass, which plays a major role in the theory.

7. THEOREM. Every point of a minimal surface $M \subset \mathbb{R}^3$ is in the image of some conformal map $\Phi: V \rightarrow M \subset \mathbb{R}^3$, where $V \subset \mathbb{C}$ is a simply-connected open set. Each such conformal map Φ is of the form $\Phi = \Phi_{(f,g)}$, where

$$\Phi_{(f,g)}^1(x, y) = \operatorname{Re} \int \frac{1}{2} f(w)(1 - g(w)^2) dw + c_1$$

$$\Phi_{(f,g)}^2(x, y) = \operatorname{Re} \int \frac{i}{2} f(w)(1 + g(w)^2) dw + c_2$$

$$\Phi_{(f,g)}^3(x, y) = \operatorname{Re} \int f(w)g(w) dw + c_3.$$

In these equations, the c_i are real numbers, g is meromorphic on V , and f is an analytic function on V vanishing precisely at the poles of g , the order of the zero being exactly twice the order of the pole; the integrals are taken along any path from a fixed point $x_0 + iy_0 \in V$ to the point $x + iy$.

Conversely, every such $\Phi_{(f,g)}$ is a conformal map into a minimal surface.

PROOF. We have already seen that there is a conformal map $\Phi: V \rightarrow M \subset \mathbb{R}^3$ given by

$$(1) \quad \Phi^k(x, y) = \operatorname{Re} \phi_k(x + iy),$$

for complex analytic functions ϕ_k satisfying

$$(2) \quad \sum_k (\phi_k')^2 = 0.$$

By Lemma 6 we have

$$(3) \quad \phi_1' = \frac{1}{2}f(1 - g^2), \quad \phi_2' = \frac{i}{2}f(1 + g^2), \quad \phi_3' = fg,$$

where f has a zero of order at least $2m$ at each point where g has a pole of order m . We just have to show that the order of f is exactly $2m$ at such a pole. Now if we had $\phi_1'(x + iy) = \phi_2'(x + iy) = 0$, then we would also have $\phi_3'(x + iy) = 0$ by (2). Since

$$(4) \quad \phi_k'(x + iy) = \frac{\partial \Phi^k}{\partial x} - i \frac{\partial \Phi^k}{\partial y},$$

this would mean that $\partial \Phi / \partial x = \partial \Phi / \partial y = 0$ at (x, y) , contradicting the fact that Φ is conformal (and hence an immersion). So $\phi_k'(x + iy) \neq 0$ for $k = 1$ or 2 (or both). Then equation (3) implies that the order of f is at most $2m$ at a pole of g of order m .

Conversely, consider $\Phi = \Phi_{(f,g)}$ where f and g have the stated properties. Then we have equation (1), where the ϕ_k are given by (3), and hence satisfy (2). It follows from (2) and (4) that

$$(5) \quad \left\langle \frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial x} \right\rangle = \left\langle \frac{\partial \Phi}{\partial y}, \frac{\partial \Phi}{\partial y} \right\rangle, \quad \left\langle \frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y} \right\rangle = 0.$$

Now our hypotheses on f and g imply [by (3)] that ϕ_1' and ϕ_2' are nowhere zero, and thus that $\partial\Phi/\partial x$ and $\partial\Phi/\partial y$ are nowhere zero. Since they are also orthogonal, by (5), they are linearly independent, so the map Φ is an immersion, and thus a conformal immersion into its image. Since the Φ^k are the real parts of complex analytic functions, they satisfy Laplace's equation, so Φ is also a minimal immersion, by Proposition 5. ♦

In order to connect this with the differential geometric properties of minimal surfaces, we need the following additional information, which will also explain the significance of the poles of g .

8. PROPOSITION. For the immersion $\Phi = \Phi_{(f,g)}$ of Theorem 7, the metric $\Phi^*\langle \cdot, \cdot \rangle$ on V has components $g_{ij} = \mu\delta_{ij}$, where

$$\mu(z) = \left[\frac{|f(z)|(1 + |g(z)|^2)}{2} \right]^2$$

[this expression will approach some limit at z if z is a pole of g].

If N is the normal map of Φ , then

$$N(z) = \left(\frac{2 \operatorname{Re} g(z)}{|g(z)|^2 + 1}, \frac{2 \operatorname{Im} g(z)}{|g(z)|^2 + 1}, \frac{|g(z)|^2 - 1}{|g(z)|^2 + 1} \right) \in S^2$$

[$= (0, 0, 1) \in S^2$ if z is a pole of g].

PROOF. Since Φ is conformal, we have $g_{ij} = \mu\delta_{ij}$, where

$$\mu = \left\langle \frac{\partial\Phi}{\partial x}, \frac{\partial\Phi}{\partial x} \right\rangle = \left\langle \frac{\partial\Phi}{\partial y}, \frac{\partial\Phi}{\partial y} \right\rangle.$$

Using

$$\phi_k'(x + iy) = \frac{\partial\Phi^k}{\partial x} - i \frac{\partial\Phi^k}{\partial y}$$

and equation (3) of the previous proof, this gives

$$\mu(z) = \frac{1}{2} \sum_k |\phi_k'(z)|^2 = \left[\frac{|f(z)|(1 + |g(z)|^2)}{2} \right]^2.$$

We also see that at points where g does not have a pole, we have

$$\begin{aligned} \frac{\partial \Phi}{\partial x} \times \frac{\partial \Phi}{\partial y} &= (\operatorname{Re} \phi_1', \operatorname{Re} \phi_2', \operatorname{Re} \phi_3') \times -(\operatorname{Im} \phi_1', \operatorname{Im} \phi_2', \operatorname{Im} \phi_3') \\ &= (\operatorname{Re} \phi_3' \operatorname{Im} \phi_2' - \operatorname{Re} \phi_2' \operatorname{Im} \phi_3', \dots) \\ &= (\operatorname{Im} \phi_2 \bar{\phi}_3, \operatorname{Im} \phi_3 \bar{\phi}_1, \operatorname{Im} \phi_1 \bar{\phi}_2) \\ &= \frac{|f|^2(1 + |g|^2)}{4} (2 \operatorname{Re} g, 2 \operatorname{Im} g, |g|^2 - 1). \end{aligned}$$

From this we compute that

$$\left| \frac{\partial \Phi}{\partial x} \times \frac{\partial \Phi}{\partial y} \right| = \left[\frac{|f|(1 + |g|^2)}{2} \right]^2 = \mu,$$

which we should have known anyway, and finally get

$$\frac{\frac{\partial \Phi}{\partial x} \times \frac{\partial \Phi}{\partial y}}{\left| \frac{\partial \Phi}{\partial x} \times \frac{\partial \Phi}{\partial y} \right|} = \left(\frac{2 \operatorname{Re} g}{|g|^2 + 1}, \frac{2 \operatorname{Im} g}{|g|^2 + 1}, \frac{|g|^2 - 1}{|g|^2 + 1} \right).$$

As we approach a pole, this clearly approaches $(0, 0, 1)$, since $g \rightarrow \infty$. ♦

The representation in Theorem 7 is not unique, because there are many different conformal maps $\Phi: V \rightarrow M$. If $\Phi_i: V_i \rightarrow M$ are two conformal maps, then the map

$$\alpha = \Phi_2^{-1} \circ \Phi_1: U \rightarrow \mathbb{R}^2 \quad U = \Phi_1^{-1}(\Phi_2(V_2)),$$

from the open set $U \subset \mathbb{R}^2$ into \mathbb{R}^2 , is conformal with respect to the usual Riemannian metric on \mathbb{R}^2 . It is easy to see (Problem 4-9) that such conformal maps α are precisely the one-one complex analytic maps α and their conjugates. Conversely, if we are given $\Phi_{(f,g)}: V \rightarrow M$ in Theorem 7, and a one-one analytic or conjugate analytic map $\alpha: W \rightarrow V$, then $\Phi_{(f,g)} \circ \alpha: W \rightarrow M$ is another conformal map into the same minimal surface, and it must have the same form, with different f and g . One can obtain the new f and g by making the substitution $w = \alpha(u)$ in the integrals of Theorem 7.

The non-uniqueness in Theorem 7 is not really much of a problem, for we have already seen that there is practically a canonical way to select a conformal map $\Phi: V \rightarrow M$ which covers a given point p of an imbedded minimal surface $M \subset \mathbb{R}^3$. We only have to assume that p is not a flat point, and also that $v(p) \neq (0, 0, 1) \in S^2$. Then $v(U) \subset S^2 - \{(0, 0, 1)\}$ for some neighborhood U of p , and $\sigma \circ v: U \rightarrow V \subset \mathbb{C}$ is conformal, where $\sigma: S^2 - \{(0, 0, 1)\} \rightarrow \mathbb{C}$ is stereographic projection. Hence we can choose $v^{-1} \circ \sigma^{-1}: V \rightarrow \mathbb{R}^3$ as our conformal map, and Theorem 7 shows that there are f and g with

$$v^{-1} \circ \sigma^{-1} = \Phi_{(f,g)}, \quad \text{or} \quad N = v \circ \Phi_{(f,g)} = \sigma^{-1}.$$

But the formula in Proposition 8, together with the formula for σ^{-1} on page 265, shows that $N = \sigma^{-1}$ precisely when $g(z) = z$ for all z . We therefore have a representation of M in the following form (traditionally written with omission of the constants c_i):

$$\begin{aligned} \Phi^1 &= \operatorname{Re} \int \frac{1}{2} F(w)(1 - w^2) dw \\ (*) \quad \Phi^2 &= \operatorname{Re} \int \frac{i}{2} F(w)(1 + w^2) dw \quad (F \text{ nowhere } 0). \\ \Phi^3 &= \operatorname{Re} \int F(w)w dw \end{aligned}$$

We could also have obtained this representation in a different way, by beginning with the formulas for $\Phi_{(f,g)}$ in Theorem 7 and then making the substitution $w = g^{-1}(u)$; in other words, we could find the formulas for $\Phi_{(f,g)} \circ g^{-1}$. Notice that a local inverse g^{-1} exists around z precisely when z is not a pole of g and $g'(z) \neq 0$; the first condition is equivalent to $v(\Phi(z)) \neq (0, 0, 1)$, and it is easy to see that the second condition is equivalent to v_* being one-one at $\Phi(z)$.

The representation $(*)$ is especially nice to work with. Problem 1 gives the choices of F which lead to the helicoid, the catenoid, and Scherk's minimal surface; if we take the simplest case of all, $F(w) = 1$, we obtain Enneper's surface, which seemed so mysterious when it was first introduced in Chapter 3. Naturally, the geometric information given by Proposition 8 now simplifies considerably. If Φ_F is given by $(*)$, then

$$\begin{aligned} N &= v \circ \Phi_F = \sigma^{-1} \\ (**) \quad \Phi_F^* \langle \cdot, \cdot \rangle &= \mu(dx \otimes dx + dy \otimes dy), \\ \text{where } \mu(z) &= \frac{|F(z)|^2(1 + |z|^2)^2}{4}. \end{aligned}$$

Notice in particular, that for real θ , the minimal surfaces $\Phi = \Phi_{e^{-i\theta}F}$,

$$\Phi^1 = \operatorname{Re} e^{-i\theta} \int \frac{1}{2} F(w)(1 - w^2) dw$$

$$\Phi^2 = \operatorname{Re} e^{-i\theta} \int \frac{i}{2} F(w)(1 + w^2) dw$$

$$\Phi^3 = \operatorname{Re} e^{-i\theta} \int F(w)w dw,$$

are all locally *isometric*, the isometry being given by

$$\Phi_{e^{-i\theta}F}(z) \mapsto \Phi_{e^{-i\phi}F}(z).$$

In general, we call two connected minimal surfaces **associated** if they have this representation for the same F and real θ and ϕ . It suffices to have this for some small piece of each surface, since minimal surfaces are analytic. We also define two planes to be associated surfaces (these are the only minimal surfaces where the representation $(*)$ cannot be achieved [except at isolated points]). Associated minimal surfaces are not only locally isometric, but can also clearly be made part of a continuous family of isometric surfaces. With the proper choice of F we obtain (Problem 1) the continuous family of isometric surfaces between the catenoid and helicoid which is pictured on pg. III.171.

On first consideration, it seems to be a pure stroke of luck that the catenoid and helicoid are not only isometric, but also associated. However, there's definitely more to it than that:

9. THEOREM (H. SCHWARZ). If two minimal surfaces are isometric, then one of them is congruent to an associated surface of the other.

PROOF. If one of the surfaces is a plane, the other must be also; for $H = 0$ and $K = 0$ implies that both principal curvatures are 0. So we will assume neither is a plane. We can then represent them as

$$\begin{aligned} f &= \Phi_F: V \rightarrow \mathbb{R}^3 \\ g &= \Phi_G: W \rightarrow \mathbb{R}^3. \end{aligned}$$

By hypothesis, there is a map $\alpha: V \rightarrow W$ such that the correspondence $\Phi_F(z) \mapsto \Phi_G(\alpha(z))$ is an isometry. We want to show that after changing the second minimal surface by a congruence we will actually have $\alpha = \text{identity}$. Then relations $(**)$ will show that $|F(z)| = |G(z)|$, and the maximum modulus principle applied to G/F will imply that we have $G = e^{-i\theta} F$ for some real θ .

The third fundamental form will play a role. Since the surfaces f and $g \circ \alpha$ are minimal, Proposition 2-6 gives

$$\begin{aligned} \text{III}_f &= -(K \circ f)I_f \\ \text{III}_{g \circ \alpha} &= -(K \circ g \circ \alpha)I_{g \circ \alpha}. \end{aligned}$$

On the other hand, $I_f = I_{g \circ \alpha}$ by hypothesis, and therefore $K \circ f = K \circ g \circ \alpha$ by the Theorema Egregium. So

$$\text{III}_f = \text{III}_{g \circ \alpha}.$$

But by Proposition 2-7 we have

$$\begin{aligned} \text{III}_f &= I_{N_1} = -\text{II}_{N_1} & N_1 &= \text{normal map of } f \\ \text{III}_{g \circ \alpha} &= I_{N_2} = -\text{II}_{N_2} & N_2 &= \text{normal map of } g \circ \alpha. \end{aligned}$$

We thus find that

$$I_{N_1} = I_{N_2} \quad \text{and} \quad \text{II}_{N_1} = \text{II}_{N_2}.$$

The Fundamental Theorem of Surface Theory then implies that N_1 and N_2 are the same up to a congruence. So if we change our second surface by a congruence we can assume that $N_1 = N_2$. But then (**) gives

$$\sigma^{-1}(z) = \sigma^{-1}(\alpha(z)).$$

So we must have $\alpha(z) = z$. ♦

We conclude with one curious phenomenon concerning the representation (*). This representation was supposed to depend only on the imbedded minimal surface M , but this is not exactly the case, for it also depends on the choice of the normal map ν , or equivalently, on the choice of an orientation for M . So while ν gives rise to the map Φ_F with

$$(1) \quad \nu \circ \Phi_F = \sigma^{-1} \quad \text{defined on some } V \subset \mathbb{C},$$

the map $-\nu$ will give rise to a map $\Phi_{\tilde{F}}$ with

$$(2) \quad -\nu \circ \Phi_{\tilde{F}} = \sigma^{-1} \quad \text{defined on some } W \subset \mathbb{C}.$$

Since Φ_F and $\Phi_{\tilde{F}}$ are conformal maps into M , inducing opposite orientations, there must be a conjugate analytic map $\alpha: W \rightarrow V$ such that

$$(3) \quad \Phi_F(z) = \Phi_{\tilde{F}}(\alpha(z)) \quad z \in W.$$

This means that for all $z \in W$ we have

$$\begin{aligned}\sigma^{-1}(\alpha(z)) &= -v(\Phi_{\tilde{F}}(\alpha(z))) && \text{by (2)} \\ &= -v(\phi_F(z)) && \text{by (3)} \\ &= -\sigma^{-1}(z) && \text{by (1).}\end{aligned}$$

Thus we must have

$$\begin{aligned}\alpha(z) &= \sigma(-\sigma^{-1}(z)) \\ &= -\frac{1}{\bar{z}}.\end{aligned}$$

Writing equation (3) in terms of (*), we thus obtain

$$\begin{aligned}\operatorname{Re} \int^z F(w)(1-w^2) dw (+ \text{constant}) &= \operatorname{Re} \int^{-1/\bar{z}} \tilde{F}(w)(1-w^2) dw \\ &= \operatorname{Re} \left(\overline{\int^{-1/\bar{z}} \tilde{F}(w)(1-w^2) dw} \right) \\ &= \operatorname{Re} \int^{-1/z} \overline{\tilde{F}(\bar{w})(1-\bar{w}^2)} dw \\ &= \operatorname{Re} \int^{-1/z} \overline{\tilde{F}(\bar{w})} (1-w^2) dw,\end{aligned}$$

which, using substitution, yields

$$\operatorname{Re} \int^z F(w)(1-w^2) dw (+ \text{constant}) = \operatorname{Re} \int^z \overline{\tilde{F}\left(-\frac{1}{\bar{w}}\right)} \left(1 - \frac{1}{w^2} \frac{1}{\bar{w}^2}\right) dw.$$

We obtain two other equations in a similar way, but, as one would certainly hope, these equations all lead to the same relation:

$$\tilde{F}(z) = -\frac{1}{z^4} \overline{F\left(-\frac{1}{\bar{z}}\right)}.$$

This \tilde{F} gives the exact same surface as F , but it induces the opposite orientation on M .

Now the interesting thing is, that there are functions F which equal \tilde{F} , the simplest example being

$$F(z) = 1 - \frac{1}{z^4} = \frac{z^4 - 1}{z^4}.$$

This choice of F leads to **Henneberg's minimal surface**

$$\Phi^1 = \operatorname{Re} \int \frac{1}{2} \frac{w^4 - 1}{w^4} (1 - w^2) dw$$

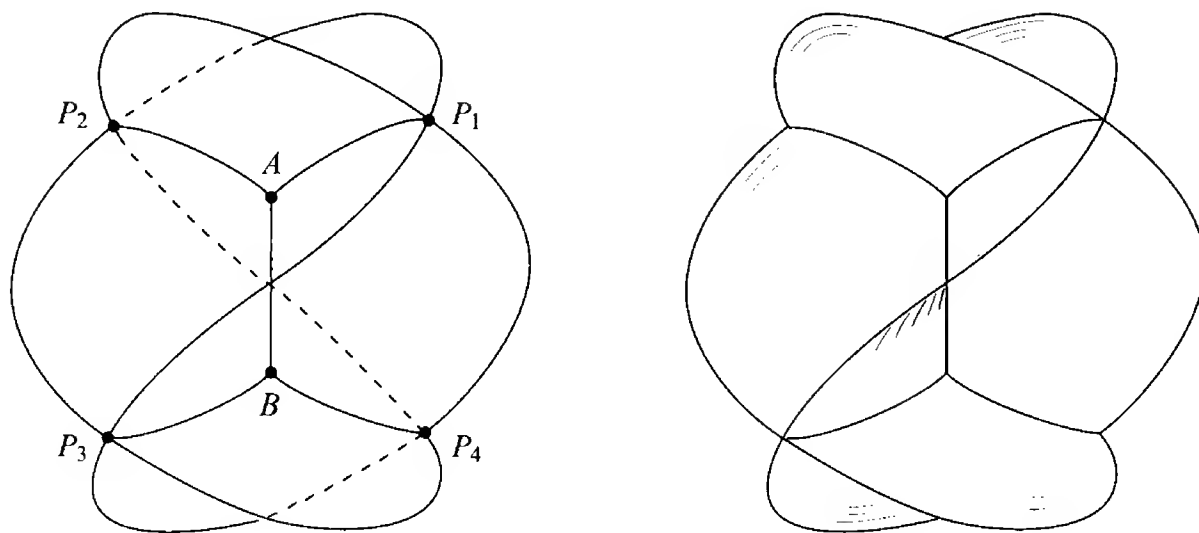
$$\Phi^2 = \operatorname{Re} \int \frac{i}{2} \frac{w^4 - 1}{w^4} (1 + w^2) dw$$

$$\Phi^3 = \operatorname{Re} \int \frac{w^4 - 1}{w^3} dw.$$

The map Φ can be defined on all of $\mathbb{C} - \{0\}$ (we don't even have to restrict ourselves to a simply-connected domain, since all integrands have residue 0 at 0, so the integrals are independent of the path); however, Φ is not an immersion at $\pm 1, \pm i$, the points where F is zero. Using stereographic projection, we can identify $\mathbb{C} - \{0, \pm 1, \pm i\}$ with S^2 minus three pairs of antipodal points, the points $\pm 1, \pm i$ occurring on the equator of S^2 . Since

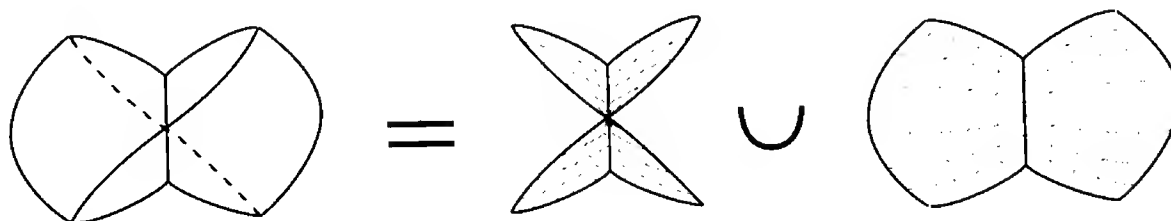
$$\Phi_F(z) = \Phi_{\tilde{F}}(\alpha(z)) = \Phi_{\tilde{F}}(\sigma(-\sigma^{-1}(z))) = \Phi_F(\sigma(-\sigma^{-1}(z))),$$

the map $\Phi_F \circ \sigma^{-1}: S^2 \rightarrow \mathbb{R}^3$ is invariant under the antipodal map, so our surface is the image of the projective plane punctured at three points. The figure below shows the image of a symmetric strip around the equator of S^2 .

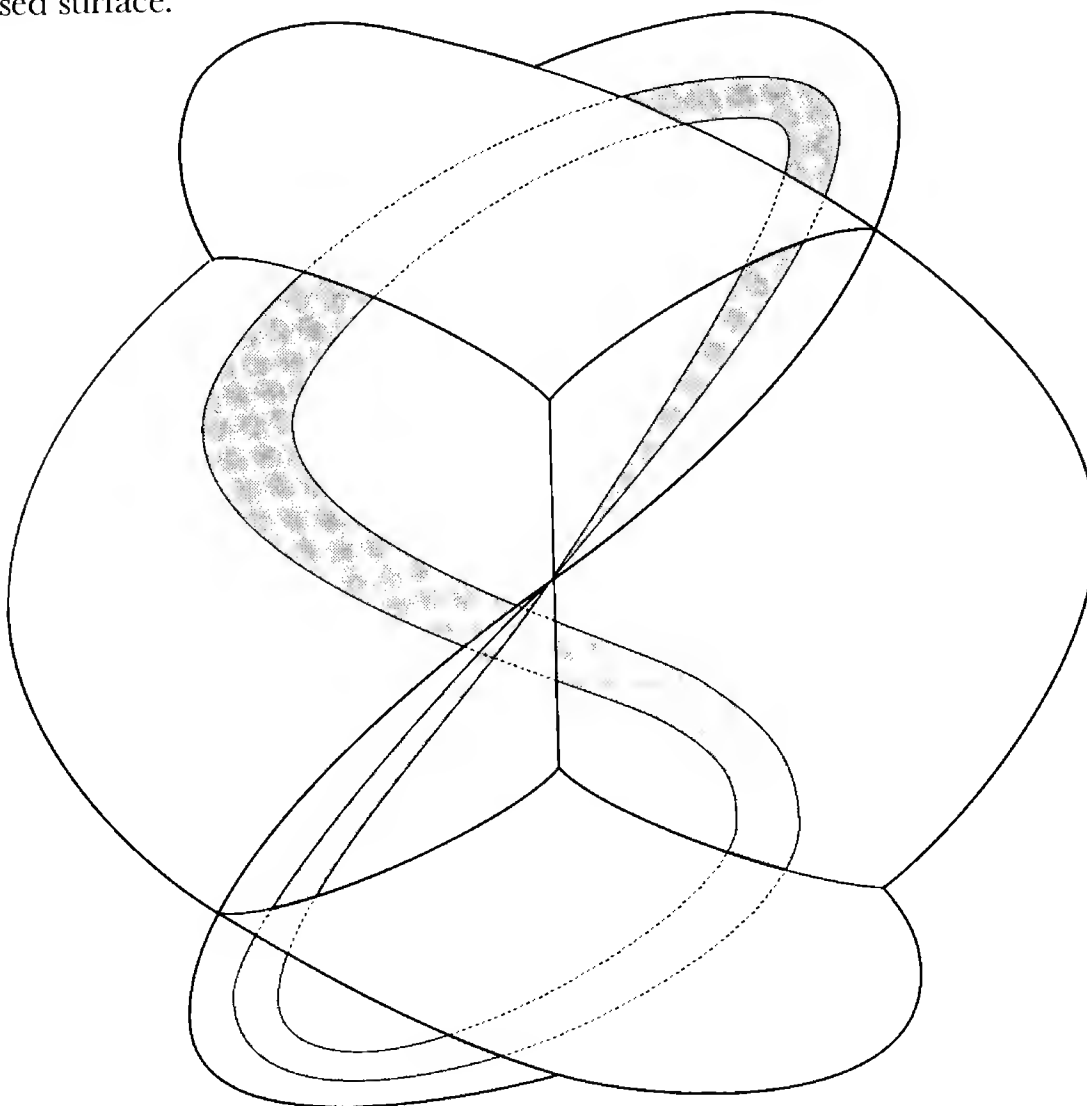


The equator maps onto the vertical segment AB , with the pair of points corresponding to $\pm i$ mapping onto the upper endpoint, and the pair ± 1 onto the lower. The boundary circles of the strip each map into the closed curve which

intersects itself at points P_1, \dots, P_4 . The points of the segment AB are all double points of the immersion, but the surface crosses itself in a funny way along this line—it contains two congruent helicoid-like surfaces, with AB common to both.

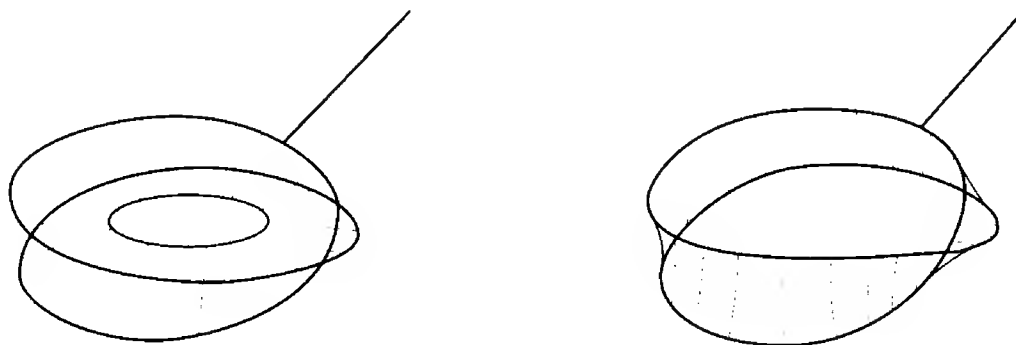


The final figure below shows an imbedded Möbius strip lying inside the immersed surface.



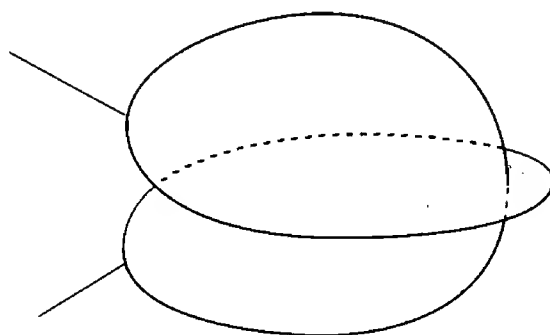
Physicists, by the way, would not be surprised to learn that there are minimal surfaces in the shape of a Möbius strip. If one dips an appropriately bent piece

of wire into a soap solution, then a soap film will be formed in the shape of this surface (actually, one always obtains the sort of film pictured on the left; after the middle sheet is pierced, the soap film snaps back into the Möbius strip). If



one neglects the slight effect of gravity, then any soap film ought to be a minimal surface, since the surface tension makes the film contract as much as possible.

Considerations of this sort were first introduced by the blind experimental physicist Plateau, who gave a much more elaborate discussion of the problem, taking into account the thickness of the films. His writings gave rise to the **Plateau problem**, to prove that every imbedded circle in \mathbb{R}^3 is the boundary of an immersed disc which has minimum area among all such immersed discs; this very difficult problem was first solved by Jesse Douglas and Tibor Rado. Douglas' methods work just as well for higher dimensions, and his work won him the Field's medal in 1936. We will not even enter into a discussion of this work, which is almost purely analytic in nature, but descriptions of the methods used may be found in several references in the bibliography. There are many questions related to Plateau's problem, some of which have led to the invention of powerful new techniques. Notice, for example, that Plateau's problem is in some ways not even the natural question to ask, since it is concerned only with surfaces homeomorphic to a disc. Thus the solution of the Plateau problem for the curve pictured above will not be the Möbius strip, but a surface like the one shown below. Probably the simplest way to picture this surface is to make a piece



of wire in the right shape and dip it into a bubble solution (the two loops should be rather further apart than in the previous picture). It is fairly easy to find a shape that gives both a Möbius strip and a disc, depending on how it is dipped in. Since the two different soap films have unequal areas, this shows us that we should slightly revise our criterion for the shape of a soap film spanned by a given wire loop. The film need not have a minimum area—a local minimum should suffice. A surface which is a critical value, but not a local minimum, would presumably correspond to a position of unstable equilibrium—the slightest disturbance would cause the soap film to change shape; presumably such films could never occur in practice (in addition, of course, all sorts of physical considerations might rule out other surfaces on practical grounds).

If the wire loop is equipped with a pair of handles, then by gently pulling the two parts of the loop apart one can see the film jump from a Möbius strip to a disc, presumably at the point where the Möbius strip is no longer in stable equilibrium. Even for those who are willing to get involved in all the analysis necessary for the Plateau problem, experiments like this can be as instructive as they are fascinating, and provide convincing evidence for assertions that are still not mathematically provable; the interested reader should consult Courant [1]. And even if you are not eager to get your hands all soaped up, there is one description of simple experiments that you simply cannot afford to miss. This is a series of lectures by Boys [1] which treats soap films and soap bubbles, the mathematical correlates of which we will study a little later on. They were given to an audience of children in the good old Victorian days, and are among the best science writing ever produced. I seriously suggest that you put down the silly stuff you are presently reading, rush right out to purchase Boys' little gem of a book, and get high on physics for a while.

* * *

Returning to purely mathematical questions, we now seek a formula for the variation of area when we are dealing with an arbitrary variation of an immersed surface $f: M \rightarrow N$, in a general Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$. We would even like to find the variation of n -dimensional volume for an immersion $f: M^n \rightarrow N^m$ (but at least we will not worry about maps and variations which are only piecewise C^∞). As a start in this direction, we consider a simple general problem from the classical calculus of variations in several variables. Suppose we are given a (suitably differentiable) function

$$F: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$$

and a compact n -dimensional manifold-with-boundary $D \subset \mathbb{R}^n$. We seek, among all functions $g: D \rightarrow \mathbb{R}$ with prescribed values on ∂D , one which will

maximize (or minimize) the quantity

$$\begin{aligned}
 J(g) &= \int_D F(t_1, \dots, t_n, g(t_1, \dots, t_n), D_1 g(t_1, \dots, t_n), \dots, D_n g(t_1, \dots, t_n)) dt_1 \dots dt_n \\
 &= \int_D F(t, g(t), Dg(t)) dt_1 \dots dt_n, \quad \text{in abbreviated form.}
 \end{aligned}$$

This is a direct generalization of the problem considered on pg. I.316. For any variation $\alpha: (-\varepsilon, \varepsilon) \times D \rightarrow \mathbb{R}$ of g , we compute the variation of J as follows. It will be convenient to denote a typical point in the domain of F by

$$(t_1, \dots, t_n, x, y_1, \dots, y_n) \quad \text{or, even more briefly, by} \quad (t, x, y).$$

Then

$$\begin{aligned}
 \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} &= \left. \frac{d}{du} \right|_{u=0} \int_D F \left(t, \alpha(u, t), \frac{\partial \alpha}{\partial t}(u, t) \right) dt_1 \dots dt_n \\
 &\quad \left(\frac{\partial \alpha}{\partial t} \text{ stands for } \frac{\partial \alpha}{\partial t_1}, \dots, \frac{\partial \alpha}{\partial t_n} \right) \\
 &= \int_D \left[\left. \frac{d}{du} \right|_{u=0} F(t, \alpha(u, t), \frac{\partial \alpha}{\partial t}(u, t)) \right] dt_1 \dots dt_n \\
 &= \int_D \left[\frac{\partial \alpha}{\partial u}(0, t) \cdot \frac{\partial F}{\partial x}(\bullet) + \sum_{i=1}^n \frac{\partial^2 \alpha}{\partial u \partial t_i}(0, t) \cdot \frac{\partial F}{\partial y_i}(\bullet) \right] dt_1 \dots dt_n,
 \end{aligned}$$

where

$$(1) \quad \bullet = \left(t_1, \dots, t_n, g(t_1, \dots, t_n), \frac{\partial g}{\partial t_1}(t_1, \dots, t_n), \dots, \frac{\partial g}{\partial t_n}(t_1, \dots, t_n) \right).$$

Introducing the abbreviations

$$\begin{aligned}
 (2) \quad w(t) &= \frac{\partial \alpha}{\partial u}(0, t) \\
 A(t) &= \frac{\partial F}{\partial x}(\bullet) \quad (\text{all of these are functions on } D) \\
 B_i(t) &= \frac{\partial F}{\partial y_i}(\bullet).
 \end{aligned}$$

we can write

$$\left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_D \left[w \cdot A + \sum_{i=1}^n \frac{\partial w}{\partial t_i} \cdot B_i \right] dt_1 \wedge \dots \wedge dt_n.$$

We now have to pull an integration-by-parts-type trick on the second term in the integrand. We do this by considering the $(n-1)$ -form ϖ defined by

$$(3) \quad \varpi = \sum_{i=1}^n (-1)^{i+1} (w \cdot B_i) dt_1 \wedge \cdots \wedge \widehat{dt_i} \wedge \cdots \wedge dt_n.$$

Since

$$d\varpi = \left[w \cdot \sum_{i=1}^n \frac{\partial B_i}{\partial t_i} \right] dt_1 \wedge \cdots \wedge dt_n + \left[\sum_{i=1}^n \frac{\partial w}{\partial t_i} B_i \right] dt_1 \wedge \cdots \wedge dt_n,$$

we have

$$\begin{aligned} \frac{dJ(\bar{\alpha}(u))}{du} \Big|_{u=0} &= \int_D \left[w \cdot \left(A - \sum_{i=1}^n \frac{\partial B_i}{\partial t_i} \right) \right] dt_1 \wedge \cdots \wedge dt_n + \int_D d\varpi \\ &= \int_D \left[w \cdot \left(A - \sum_{i=1}^n \frac{\partial B_i}{\partial t_i} \right) \right] dt_1 \wedge \cdots \wedge dt_n + \int_{\partial D} \varpi. \end{aligned}$$

From the definition of ϖ , we see that $\varpi = 0$ on ∂D if α is a variation keeping the boundary fixed. So g is a critical point for J if and only if

$$0 = \frac{dJ(\bar{\alpha}(u))}{du} \Big|_{u=0} = \int_D \left[\frac{\partial \alpha}{\partial u}(0, t) \cdot \left(A - \sum_{i=1}^n \frac{\partial B_i}{\partial t_i} \right) \right] dt_1 \wedge \cdots \wedge dt_n$$

for all variations α keeping the boundary fixed. From this we easily see that g must satisfy the equation

$$A - \sum_{i=1}^n \frac{\partial B_i}{\partial t_i} = 0,$$

that is,

$$(*) \quad \frac{\partial F}{\partial x}(\bullet) - \sum_{i=1}^n \frac{\partial^2 F}{\partial t_i \partial y_i}(\bullet) = 0, \quad \text{where } \bullet \text{ is given by (1).}$$

This is the classical analogue of Euler's equation (Theorem I.9-8).

As a particular example, we take $n = 2$, and let

$$(4) \quad F(t_1, t_2, x, y_1, y_2) = \sqrt{1 + y_1^2 + y_2^2}.$$

so that

$$\begin{aligned} J(g) &= \int_D \sqrt{1 + g_1^2 + g_2^2} \, dt_1 \, dt_2 \quad \left(g_i = \frac{\partial g}{\partial t_i} \right) \\ &= \int_D \sqrt{EG - F^2} \, dt_1 \, dt_2 \quad \text{by formulas (A') on pg. III.137.} \end{aligned}$$

Thus $J(g)$ is the area of the imbedded surface

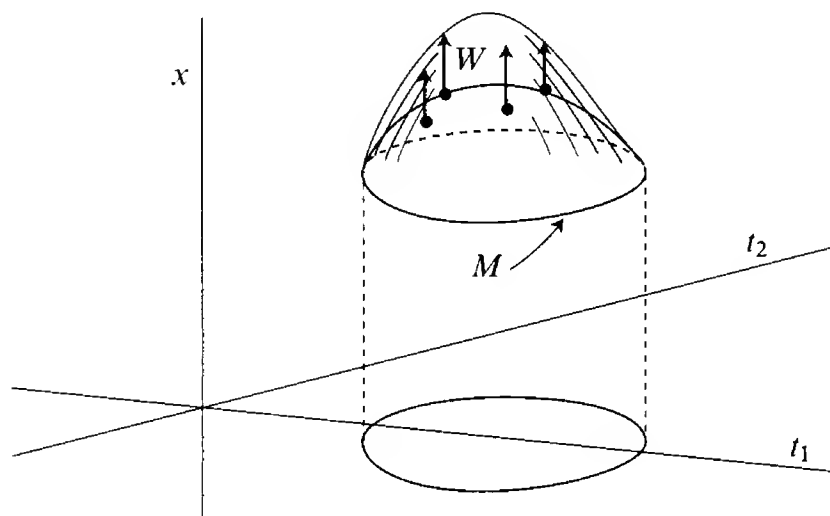
$$f(t_1, t_2) = (t_1, t_2, g(t_1, t_2)).$$

A variation $\alpha: (-\varepsilon, \varepsilon) \times D \rightarrow \mathbb{R}$ of g gives rise to a variation $\beta: (-\varepsilon, \varepsilon) \times D \rightarrow \mathbb{R}^3$ of the imbedding f , defined by

$$\beta(u, t_1, t_2) = (t_1, t_2, \alpha(u, t_1, t_2)).$$

This variation β is perpendicular to the (t_1, t_2) -plane, instead of being perpendicular to the surface $M = f(D)$; it has variation vector

$$W(t_1, t_2) = (0, 0, w(t_1, t_2))_{f(t_1, t_2)}.$$



This is the one other kind of variation sometimes encountered in differential geometry books, and the kind which is always used in books on the calculus of variations. Indeed, this particular example was chosen by Lagrange to illustrate the general methods which he had developed (1760) for the calculus of variations in several variables. In this case, equation (4) gives

$$\frac{\partial F}{\partial x} = 0, \quad \frac{\partial F}{\partial y_i} = \frac{y_i}{\sqrt{1 + y_1^2 + y_2^2}}.$$

so equation (*) becomes

$$\frac{\partial}{\partial t_1} \left(\frac{g_1}{R} \right) + \frac{\partial}{\partial t_2} \left(\frac{g_2}{R} \right) = 0, \quad R = \sqrt{1 + g_1^2 + g_2^2},$$

which boils down to exactly the equation

$$(1 + g_1^2)g_{11} - 2g_1g_2g_{12} + (1 + g_2^2)g_{22} = 0$$

which we found in the proof of Proposition 3; it was only in 1776 that Meusnier interpreted this equation in terms of the mean curvature of f .

We will also be interested in the 1-form ϖ which we obtain in this case; from (2) and (3) we see that

$$\varpi = \frac{w \cdot g_1}{R} dt_2 - \frac{w \cdot g_2}{R} dt_1.$$

We will express ϖ in terms of the form ω on $M = f(D)$ with $\varpi = f^*\omega$. We have

$$\begin{aligned} \omega((1, 0, g_1)) &= \varpi \left(f_* \left(\frac{\partial}{\partial t_1} \right) \right) = \left\langle (1, 0, g_1), \left(-\frac{w \cdot g_2}{R}, -\frac{w \cdot g_1}{R}, 0 \right) \right\rangle \\ \omega((0, 1, g_2)) &= \varpi \left(f_* \left(\frac{\partial}{\partial t_2} \right) \right) = \left\langle (0, 1, g_2), \left(-\frac{w \cdot g_2}{R}, -\frac{w \cdot g_1}{R}, 0 \right) \right\rangle. \end{aligned}$$

But

$$\begin{aligned} \left(-\frac{w \cdot g_2}{R}, -\frac{w \cdot g_1}{R}, 0 \right) &= \left(-\frac{g_1}{R}, -\frac{g_2}{R}, \frac{1}{R} \right) \times (0, 0, w) \\ &= v \times W, \end{aligned}$$

where v is the normal vector. So for all $X \in M_p$ we have

$$\begin{aligned} \omega(p)(X) &= \langle v(p) \times W(p), X \rangle \\ &= \langle W(p) \times X, v(p) \rangle \\ &= \langle \mathbb{T}W(p) \times X, v(p) \rangle, \quad \mathbb{T}W(p) = \text{tangential component of } W(p). \end{aligned}$$

If dA is the 2-dimensional volume form on M , then we have

$$\omega(X) = dA(\mathbb{T}W, X).$$

Using the notation introduced on pg. I.227, we can thus write

$$\omega = \mathbb{T}W \lrcorner dA.$$

Without going through the calculations, we merely state that if we take an arbitrary n and let

$$F(t_1, \dots, t_n, x, y_1, \dots, y_n) = \sqrt{1 + \sum_i y_i^2},$$

so that $J(g)$ represents the n -dimensional volume of the imbedded n -manifold $\{f(t_1, \dots, t_n, g(t_1, \dots, t_n))\}$, then the $(n-1)$ -form ϖ is $f^*\omega$, where the $(n-1)$ -form ω on M is defined by

$$\omega = \mathbb{T}W \lrcorner dV \quad dV = \text{volume element on } M.$$

This $(n-1)$ -form ω will be very important when we look for an invariant description of the variation of n -dimensional volume for an immersion $f: M^n \rightarrow N$. We have always expressed length or area as an integral involving a coordinate system, and calculated the derivative with respect to the variation parameter u by using “Leibniz’ Rule” to bring the derivative inside the integral sign. Before we go any further, we will need an invariant description of this procedure.

Suppose we have a C^∞ 1-parameter family of k -forms on an n -manifold (-with-boundary) M ; thus, for each $u \in (-\varepsilon, \varepsilon)$, we have a k -form $\Gamma(u)$ on M . For each $p \in M$, the map $u \mapsto \Gamma(u)(p) \in \Omega^k(M_p)$ into the vector space $\Omega^k(M_p)$ then has a derivative, which at each u is again an element $\dot{\Gamma}(u)(p) \in \Omega^k(M_p)$. Thus a C^∞ 1-parameter family of k -forms $u \mapsto \Gamma(u)$ on M gives rise to a new C^∞ 1-parameter family of k -forms $u \mapsto \dot{\Gamma}(u)$ on M .

10. PROPOSITION (LEIBNIZ’ RULE). Let M be a compact oriented n -dimensional manifold-with-boundary and $u \mapsto \Gamma(u)$ a C^∞ 1-parameter family of n -forms on M . Then

$$\left. \frac{d}{du} \right|_{u=u_0} \int_M \Gamma(u) = \int_M \dot{\Gamma}(u_0).$$

PROOF. Let \mathcal{O} be a finite cover of M by open sets V each contained in $c([0, 1]^n)$ for some orientation preserving singular n -cube $c: [0, 1]^n \rightarrow M$. Let $\Phi = \{\phi_V\}$ be a partition of unity subordinate to this cover. Then

$$\int_M \phi_V \cdot \Gamma(u) = \int_{[0, 1]^n} (\phi_V \circ c) \cdot c^* \Gamma(u).$$

It is easy to see that the ordinary Leibniz' Rule implies that

$$\left. \frac{d}{du} \right|_{u=u_0} \int_M \phi_V \cdot \Gamma(u) = \int_{[0,1]^n} (\phi_V \circ c) \cdot c^* \dot{\Gamma}(u_0) = \int_M \phi_V \dot{\Gamma}(u_0).$$

Since

$$\int_M \Gamma(u) = \sum_{\{\phi_V\}} \int_M \phi_V \cdot \Gamma(u),$$

and similarly for $\int_M \dot{\Gamma}(u_0)$, the result follows. \blacklozenge

Now consider a compact oriented n -dimensional manifold-with-boundary M , and a C^∞ map $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$, where $(N, \langle \cdot, \cdot \rangle)$ is a Riemannian manifold. We will assume that each $\bar{\alpha}(u): M \rightarrow N$ is an immersion. Then the metric $\bar{\alpha}(u)^* \langle \cdot, \cdot \rangle$ on M determines a volume element $\Gamma(u)$ on M ; using the given orientation of M , we can consider this to be an n -form on M , which we call the **volume form**. What we want to determine is

$$\left. \frac{d}{du} \right|_{u=0} \int_M \Gamma(u).$$

According to Proposition 10, it suffices to determine $\dot{\Gamma}(0)$. For this we do not even need M to be compact.

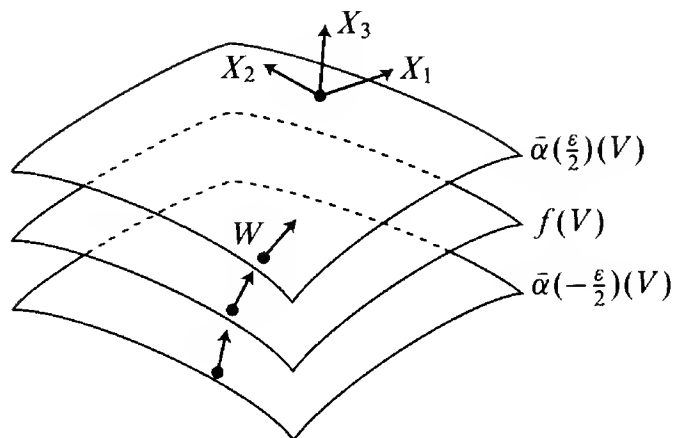
11. THEOREM (VARIATION OF VOLUME FORMULA). Let $f: M \rightarrow N$ be an immersion of an oriented n -dimensional manifold (-with-boundary) M into a Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$ and let $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$ be a variation of f through immersions, with variation vector field W . If $\Gamma(u)$ is the volume form of M determined by the metric $\bar{\alpha}(u)^* \langle \cdot, \cdot \rangle$ and the given orientation of M , then

$$\dot{\Gamma}(0) = -\langle W, n \cdot \eta \rangle \cdot \Gamma(0) + d(\top W \lrcorner \Gamma(0))$$

where η is the mean curvature normal of the immersion f . [Notice that there is a slight abuse of notation here; at each $p \in M$, the vector $\top W$ really denotes the unique vector $X \in M_p$ with $f_*(X) = \top W$ at $f(p)$.]

PROOF. The theorem involves two n -forms on M which we have to prove are equal at all points of M . Let us first consider a point $p_0 \in M$ where $W(p_0)$ is not tangent to $f(M)$. By choosing a sufficiently small neighborhood V of p_0 ,

and decreasing ε if necessary, we can then assume that $\alpha: (-\varepsilon, \varepsilon) \times V \rightarrow N$ is an imbedding.



It will be convenient to identify V with $f(V)$, so that $f = \bar{\alpha}(0)$ is just the inclusion map $i: V \rightarrow N$. On some open set containing image α , we can choose an orthonormal moving frame $X_1, \dots, X_n, X_{n+1}, \dots, X_m$ such that

- (1) $X_j(\alpha(u, p))$ is tangent to the submanifold $\bar{\alpha}(u)(V)$ $1 \leq j \leq n$
- (2) $X_r(\alpha(u, p))$ is normal to the submanifold $\bar{\alpha}(u)(V)$ $n+1 \leq r \leq m$.

If $\phi^1, \dots, \phi^n, \phi^{n+1}, \dots, \phi^m$ are the dual 1-forms, then clearly

- (1') $\bar{\alpha}(u)^*(\phi^1 \wedge \dots \wedge \phi^n) = \Gamma(u)$
- (2') $\bar{\alpha}(u)^*(\phi^r) = 0 \quad n+1 \leq r \leq m$.

Now the variation vector field W , defined along V , is the restriction of the vector field $\tilde{W} = \partial\alpha/\partial u$ defined along all of image α . We can further extend \tilde{W} to a vector field defined on some open set containing image α ; we will use the same symbol \tilde{W} for this extension. Associated to this vector field \tilde{W} is a certain local 1-parameter group of local diffeomorphisms $\{\rho_u\}$; recall (Chapter I.5) that $\rho_u(q)$ is the result of following for time u the integral curve of \tilde{W} that starts at q . Clearly the integral curve of \tilde{W} that starts at a point $p \in V$ is just $u \mapsto \alpha(u, p)$. So

$$\rho_u(p) = \alpha(u, p) = \bar{\alpha}(u)(p), \quad p \in V.$$

It is therefore clear that if Y is a tangent vector of V , then

$$(3) \quad \rho_{u*}(i_*Y) = \bar{\alpha}(u)_*(Y).$$

Now let us recall the Lie derivative (pp. I.150, 174, 234): if ω is a k -form on N , then $L_{\tilde{W}}\omega$ is another k -form defined by

$$L_{\tilde{W}}\omega(Z_1, \dots, Z_k) = \lim_{h \rightarrow 0} \frac{1}{h} [\omega(\rho_{h*}Z_1, \dots, \rho_{h*}Z_k) - \omega(Z_1, \dots, Z_k)].$$

We claim that

$$(4) \quad \dot{\Gamma}(0) = i^*\{L_{\tilde{W}}(\phi^1 \wedge \dots \wedge \phi^n)\}.$$

The proof of this will be quite straightforward. We adopt the abbreviation $\Phi = \phi^1 \wedge \dots \wedge \phi^n$. If Y_1, \dots, Y_n are tangent vectors of V , then we have

$$\begin{aligned} \dot{\Gamma}(0)(Y_1, \dots, Y_n) &= \lim_{h \rightarrow 0} \frac{1}{h} [\Gamma(h)(Y_1, \dots, Y_n) - \Gamma(0)(Y_1, \dots, Y_n)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [\bar{\alpha}(h)^*\Phi(Y_1, \dots, Y_n) - i^*\Phi(Y_1, \dots, Y_n)] \quad \text{by (1')} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [\Phi(\bar{\alpha}(h)_*Y_1, \dots, \bar{\alpha}(h)_*Y_n) - \Phi(i_*Y_1, \dots, i_*Y_n)] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [\Phi(\rho_{h*}i_*Y_1, \dots, \rho_{h*}i_*Y_n) - \Phi(i_*Y_1, \dots, i_*Y_n)] \quad \text{by (3)} \\ &= L_{\tilde{W}}\Phi(i_*Y_1, \dots, i_*Y_n), \end{aligned}$$

which proves (4).

The reason for bringing in the Lie derivative is that we have some useful formulas for it. In particular (pg. I.235), we have

$$L_{\tilde{W}}\omega = \tilde{W} \lrcorner d\omega + d(\tilde{W} \lrcorner \omega).$$

Substituting this into (4) we obtain

$$(5) \quad \dot{\Gamma}(0) = i^*\{\tilde{W} \lrcorner d\Phi\} + d(i^*\{\tilde{W} \lrcorner \Phi\}).$$

We will show that the two terms on the right are precisely the terms appearing in the statement of the theorem.

We first compute $d\Phi = d(\phi^1 \wedge \dots \wedge \phi^n)$ by using the first structural equation for N , which will bring in the connection forms ψ_β^α ($1 \leq \alpha, \beta \leq m$) for N associated to ϕ^1, \dots, ϕ^m :

$$\begin{aligned} (6) \quad d\Phi &= d(\phi^1 \wedge \dots \wedge \phi^n) = \sum_{j=1}^n (-1)^{j+1} \phi^1 \wedge \dots \wedge d\phi^j \wedge \dots \wedge \phi^n \\ &= \sum_{j=1}^n (-1)^{j+1} \phi^1 \wedge \dots \wedge \left(- \sum_{\alpha=1}^m \psi_\alpha^j \wedge \phi^\alpha \right) \wedge \dots \wedge \phi^n \\ &= \sum_{j=1}^n \sum_{r=n+1}^m \phi^r \wedge \phi^1 \wedge \dots \wedge \psi_r^j \wedge \dots \wedge \phi^n. \end{aligned}$$

So if Y_1, \dots, Y_n are the tangent vectors of V with $i_* Y_j = X_j$ along V , then

$$\begin{aligned}
 (7) \quad i^*\{\tilde{W} \lrcorner d\Phi\}(Y_1, \dots, Y_n) &= d\Phi(W, X_1, \dots, X_n) \\
 &= \sum_{j=1}^n \sum_{r=n+1}^m (\phi^r \wedge \phi^1 \wedge \dots \wedge \psi_r^j \wedge \dots \wedge \phi^n)(W, X_1, \dots, X_n) \\
 &= \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) \psi_r^j(X_j),
 \end{aligned}$$

since (2') says that $\phi^r(X_j) = 0$ for $i \leq n < r$. On the other hand, we have

$$\nabla'_{X_j} X_j = \sum_{\alpha=1}^m \psi_j^\alpha(X_j) \cdot X_\alpha = - \sum_{\alpha=1}^m \psi_\alpha^j(X_j) \cdot X_\alpha,$$

so

$$n \cdot \eta = \perp \left(\sum_{j=1}^n \nabla'_{X_j} X_j \right) = \sum_{j=1}^n \left(- \sum_{r=n+1}^m \psi_r^j(X_j) X_r \right),$$

and hence

$$(8) \quad -\langle W, n \cdot \eta \rangle = \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) \psi_r^j(X_j).$$

Equations (7) and (8) thus give

$$(9) \quad i^*\{\tilde{W} \lrcorner d\Phi\} = -\langle W, n \cdot \eta \rangle \cdot \Gamma(0).$$

As for the other term in (5), if Y_1, \dots, Y_{n-1} are tangent vectors of V , then we have

$$\begin{aligned}
 i^*\{\tilde{W} \lrcorner \Phi\}(Y_1, \dots, Y_{n-1}) &= \Phi(W, i_* Y_1, \dots, i_* Y_{n-1}) \\
 &= (\phi^1 \wedge \dots \wedge \phi^n)(W, i_* Y_1, \dots, i_* Y_{n-1}) \\
 &= (\phi^1 \wedge \dots \wedge \phi^n)(\mathbb{T}W, i_* Y_1, \dots, i_* Y_{n-1}) \\
 &\quad \text{since each } \phi^j(\perp W) = 0 \\
 &= [\mathbb{T}W \lrcorner i^*(\phi^1 \wedge \dots \wedge \phi^n)](Y_1, \dots, Y_{n-1}).
 \end{aligned}$$

Thus

$$(10) \quad i^*\{\tilde{W} \lrcorner \Phi\} = \mathbb{T}W \lrcorner \Gamma(0).$$

This completes the proof of the theorem at any point p_0 where $W(p_0)$ is not tangent to $f(M)$.

The general case can be disposed of by a technical trick. Let $\mathbf{N} = N \times \mathbb{R}$, with the product Riemannian metric, which we also denote by $\langle \cdot, \cdot \rangle$, and define $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow \mathbf{N}$ by

$$\alpha(u, p) = (\alpha(u, p), u).$$

The new variation vector field \mathbf{W} is

$$\mathbf{W}(p) = (W(p), 1),$$

where 1 denotes the unit vector field on \mathbb{R} . Clearly \mathbf{W} is not tangent to $\bar{\alpha}(0)(M) \subset N \times \{0\}$, so the theorem holds for α . On the other hand, it is easy to see that the new mean curvature normal η is just

$$\eta(p) = (\eta, 0),$$

so that $\langle \mathbf{W}, \eta \rangle = \langle W, \eta \rangle$; thus the result for α implies the result for α . ♦

12. COROLLARY. Let $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$ be a variation of an immersion $f: M \rightarrow N$ of a compact oriented n -dimensional manifold-with-boundary M into a Riemannian manifold $(N, \langle \cdot, \cdot \rangle)$. If $V(\bar{\alpha}(u))$ is the n -dimensional volume of M determined by the metric $\bar{\alpha}(u)^*\langle \cdot, \cdot \rangle$ and the given orientation of M , then

$$\left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_M \langle W, n \cdot \eta \rangle dV + \int_{\partial M} \omega,$$

where dV is the volume element determined by $f^*\langle \cdot, \cdot \rangle$ and $\omega = W \lrcorner dV$. In particular, if α is a variation keeping ∂M fixed, then

$$\left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_M \langle W, n \cdot \eta \rangle dV.$$

The immersion f is a critical point for V , among all immersions $g: M \rightarrow N$ with $g = f$ on ∂M , if and only if $\eta = 0$ everywhere.

PROOF. The first statement follows from Theorem 11, Leibniz' Rule, and Stokes' Theorem. If α keeps ∂M fixed, then $W = 0$ on ∂M , so also $\omega = 0$ on ∂M ; this proves the second statement. To prove the third, we can choose $W = \phi \cdot \eta$, where ϕ is a C^∞ function on M which is 0 on ∂M and positive on $M - \partial M$. ♦

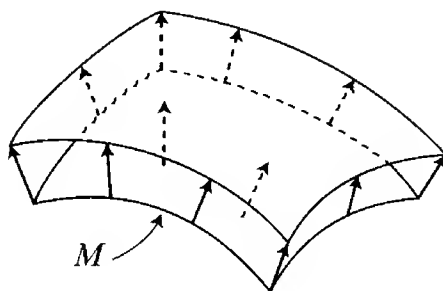
Notice that in the expression

$$- \int_M \langle W, n \cdot \eta \rangle dV + \int_{\partial M} \omega,$$

the first term depends only on the normal component $\mathbf{L}W$ of W ; for we have $\langle W, n \cdot \eta \rangle = \langle \mathbf{L}W, n \cdot \eta \rangle$, since η is perpendicular to $f(M)$. This partially confirms our suspicion that we need work only with normal variations. On the other hand, in the term $\int_{\partial M} \omega$, only the tangential component $\mathbf{T}W$ enters; roughly speaking, the integral measures how much the volume of M is changing because of the way that the variation is expanding its boundary. In particular, we see that $\int_{\partial M} \omega$ is 0 not only when the variation keeps the boundary fixed, but also when W is normal to M along the boundary. Consequently, if $\eta = 0$ on M , then $dV(\bar{\alpha}(u))/du|_{u=0}$ will be 0 for every variation which is perpendicular on the boundary of M , not merely for those variations which keep ∂M fixed. Back in our original equation (*) on page 262 we didn't have any term involving an integral over ∂D precisely because we were dealing only with normal variations. This leads to an interesting phenomenon in the case of minimal surfaces $M \subset \mathbb{R}^3$. If v is the unit normal vector on M , then we can define a variation α of the inclusion $i: M \rightarrow \mathbb{R}^3$ by

$$\alpha(u, p) = p + u \cdot v(p).$$

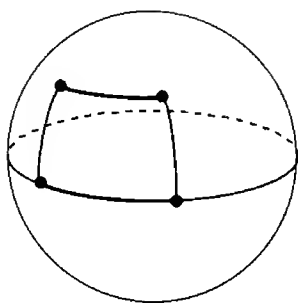
The various surfaces $\{\alpha(u, p) : p \in M\}$ are called the **parallel surfaces** of M .



Since this variation α has $W = v$, which is everywhere normal to M , we must have

$$\left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = 0.$$

But this equality does not necessarily mean that $A(\bar{\alpha}(0))$ is a minimum. Indeed, as Problem 3-12 shows, each parallel surface has *smaller* area than M , so actually $A(\bar{\alpha}(0))$ is a maximum! Something quite similar happens in the case of geodesics on a surface of positive curvature. For example, on S^2 , a portion



of a great circle is *longer* than a “parallel” curve. The phenomenon for minimum surfaces is analyzed in greater detail in Addendum 4, which considers the second variation of volume.

Although we have derived the fundamental formula for the variation of volume in all dimensions, we will not proceed to discuss the analogues of minimal surfaces in higher dimensions, except to say that this topic has generated much interest in recent years. We should also mention that the study of minimal hypersurfaces in spheres has also attracted much attention, and differs greatly from the theory for Euclidean spaces. For example (Lawson [I]), every compact orientable surface can be imbedded as a minimal surface in S^3 .

For the remainder of this chapter, we will discuss a few other topics involving the variation of volume. We will often digress quite a bit from purely differential-geometric matters, and unfortunately our remarks will not form a coherent subject like the study of minimal surfaces.

Two special cases of Corollary 12 will form the starting point of our considerations. Suppose first that M is simply a compact manifold with no boundary. Then we have

$$(I) \quad \left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_M \langle W, n \cdot \eta \rangle dV.$$

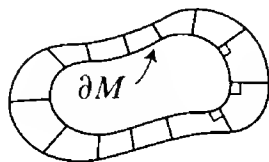
We can also apply Corollary 12 when M and N have the same dimension n , so that $M \subset N$ is a compact n -dimensional manifold-with-boundary in the n -dimensional manifold N . In this case, $M_p = N_p$ for all $p \in M$, so $\mathbf{T}: N_p \rightarrow M_p$ is the identity, while $\mathbf{\perp}: N_p \rightarrow N_p$ is the 0 map. Consequently, η is automatically 0, and we have only the boundary term left,

$$\left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_{\partial M} \omega.$$

It is easily checked that this can be written

$$(II) \quad \left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_{\partial M} \langle W, \nu \rangle dV_{n-1},$$

where ν is the outward pointing normal on ∂M and dV_{n-1} is the $(n-1)$ -dimensional volume element on ∂M . This formula is certainly reasonable, for when we move each point p on ∂M a distance $\phi(p)$ along $\nu(p)$, we add on a narrow band whose volume is approximately $\int_{\partial M} \phi dV_{n-1}$.



Both formulas (I) and (II) are important for a discussion of the **isoperimetric problem**. The classical isoperimetric problem was to find the curve of fixed length L which encloses the largest area; one naturally expects the answer to be a circle. One can also seek the curve of smallest length which encloses a fixed area; presumably the answer to this “dual” problem is also a circle. We should also mention the **problem of Dido**, to find the curve of fixed length between two points P and Q which, together with the straight line between P and Q , encloses the largest area; the expected answer is an arc of a circle. These classical problems have given rise to a whole class of problems in the calculus of variations, known generically as “isoperimetric problems”. To illustrate this sort of problem we will, for simplicity, stay in dimension 1. Consider two functions

$$F: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \quad \text{and} \quad G: \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}.$$

For a function $f: [a, b] \rightarrow \mathbb{R}$ we define

$$J(f) = \int_a^b F(t, f(t), f'(t)) dt$$

$$K(f) = \int_a^b G(t, f(t), f'(t)) dt.$$

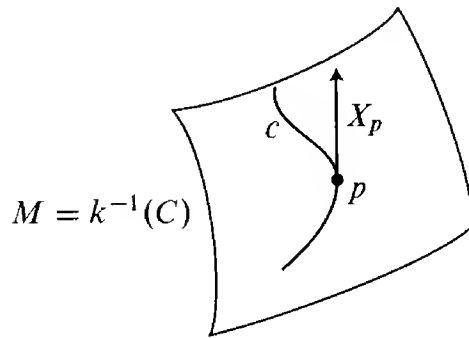
Among all functions $f: [a, b] \rightarrow \mathbb{R}$ with fixed values at a and b , and a fixed value $J(f) = C$, we seek the one which maximizes or minimizes $K(f)$. The “dual” problem is to find that f with fixed values at a and b , and fixed $K(f) = C'$, which minimizes or maximizes $J(f)$. This problem is approached by generalizing the methods which work for the corresponding problem in ordinary calculus, a review of which is now in order.

Suppose we are given two differentiable functions $j, k: \mathbb{R}^n \rightarrow \mathbb{R}$, and we seek the maximum or minimum of j on the set $k^{-1}(C)$. The method of “Lagrangian

multipliers" states that if j attains its maximum or minimum on $k^{-1}(C)$ at the point p , and p is not a critical point of k , then there is a number λ such that

$$(1) \quad \frac{\partial j}{\partial x_i}(p) = \lambda \frac{\partial k}{\partial x_i}(p) \quad i = 1, \dots, n.$$

The proof of this assertion has already been outlined in Problem 3-3, but it is so crucial to the present discussion that it will be repeated here. We note that the hypotheses on k imply that in a neighborhood of p , the set $k^{-1}(C) \subset \mathbb{R}^n$ is a hypersurface M , and that $k_*(X_p) = 0$ for $X_p \in \mathbb{R}^n_p$ precisely when $X_p \in M_p$. Every such X_p is $c'(0)$ for some curve c in M . It follows that $j(c(t))$ has a



maximum or minimum at $t = 0$, which means that $j_*(X_p) = 0$. Thus the two linear functions $j_*, k_*: \mathbb{R}^n_p \rightarrow \mathbb{R}$ have the property that $\ker k_* \subset \ker j_*$. This implies that $j_* = \lambda k_*$ for some λ , which is equivalent to equation (1).

Notice that if k attains its maximum or minimum on $g^{-1}(C')$ at q , and q is not a critical point of j , then there is a number μ such that

$$(2) \quad \frac{\partial k}{\partial x_i}(q) = \mu \frac{\partial j}{\partial x_i}(q).$$

Equations (1) and (2) are equivalent, since $\lambda, \mu \neq 0$ (as p and q are not critical points). Thus, if p is a maximum point of g on $k^{-1}(C)$ and we set $C' = j(p)$, then p is at least one of the candidates for the minimum point of k on $j^{-1}(C')$. If we simply look for critical points for j on $k^{-1}(C)$ and for k on $j^{-1}(C')$, then these two "dual" problems are completely equivalent.

Let us apply these ideas to our two functions J and K . Suppose that the maximum or minimum of J on $K^{-1}(C)$ occurs at a C^2 function f which is not a critical point of K . Consider any variation $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow \mathbb{R}$ of f which keeps endpoints fixed. We know from formula (**) on pg. I.319 that $dJ(\bar{\alpha}(u))/du|_{u=0}$ depends only on the function $\partial\alpha/\partial u(0, t)$ on $[a, b]$. For any C^2 function W on $[a, b]$ with $W(a) = W(b) = 0$, we define

$$J_{f*}(W) = \frac{dJ(\bar{\alpha}(u))}{du} \Big|_{u=0} \quad \begin{array}{l} \text{for any variation } \alpha \text{ of } f \\ \text{with } \partial\alpha/\partial u(0, t) = W(t). \end{array}$$

Notice that there always is a variation α with this property, for example,

$$\alpha(u, t) = f(t) + uW(t).$$

The same W can be used to give a variation of any function f , so the “ f ” in the symbol $J_{f*}(W)$ is important. Nevertheless, since we will be considering only one f , we will usually write simply $J_*(W)$ for convenience. We define $K_*(W)$ in precisely the same way. We thus have functions $J_*, K_*: \mathcal{V} \rightarrow \mathbb{R}$, where \mathcal{V} is the vector space of all C^2 functions W on $[a, b]$ with $W(a) = W(b) = 0$. We claim that J_* (and likewise K_*) is linear. To see this we choose two variations α_1 and α_2 with

$$\frac{\partial \alpha_i}{\partial u}(0, t) = W_i(t),$$

and define the variation α by

$$\alpha(u, t) = \alpha_1(u, t) + \alpha_2(u, t).$$

Then

$$\frac{\partial \alpha}{\partial u}(0, t) = W_1(t) + W_2(t),$$

so

$$\begin{aligned} J_*(W_1 + W_2) &= \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} \\ &= \left. \frac{dJ(\bar{\alpha}_1(u))}{du} \right|_{u=0} + \left. \frac{dJ(\bar{\alpha}_2(u))}{du} \right|_{u=0}, \\ &\quad \text{as one sees by inspecting formula (**) on pg. I.319,} \\ &= J_*(W_1) + J_*(W_2). \end{aligned}$$

Homogeneity is proved similarly.

We now make the following

CLAIM. If $K_*(W) = 0$, then $W = \partial \alpha / \partial u(0, t)$ for some variation α with the property that each $\bar{\alpha}(u)$ is in $K^{-1}(C)$.

Remember that, by hypothesis, f is *not* a critical point of K . From a modern point of view, our claim seems especially reasonable, for the set of all C^2 functions $\phi: [a, b] \rightarrow \mathbb{R}$, with given values at a and b , forms an infinite dimensional manifold, and in a neighborhood of f the set $K^{-1}(C)$ should be a submanifold of codimension 1; each “tangent vector” W at f with $K_*(W) = 0$ is a tangent vector to the submanifold $K^{-1}(C)$ and should therefore come from a “curve” α in $K^{-1}(C)$. The classical argument runs as follows.

13. LEMMA. If $K(f) = C$, where the C^2 function f is not a critical point of K , and $K_*(W) = K_{f*}(W) = 0$, then $W = \partial\alpha/\partial u(0, t)$ for some variation α with the property that each $\bar{\alpha}(u)$ is in $K^{-1}(C)$.

PROOF. Since f is not a critical point, there is W_1 with $K_*(W_1) \neq 0$. Let $L: \mathbb{R}^2 \rightarrow \mathbb{R}$ be

$$L(r, s) = K(f + rW + sW_1).$$

If we define

$$\beta(u, t) = f(t) + uW_1(t),$$

then β is a variation of f with $\partial\beta/\partial u(0, t) = W_1(t)$ and $\bar{\beta}(u) = f + uW_1$. So

$$K_*(W_1) = \lim_{u \rightarrow 0} \frac{K(f + uW_1) - K(f)}{u} = \frac{\partial L}{\partial s}(0, 0).$$

Similarly,

$$K_*(W) = \frac{\partial L}{\partial r}(0, 0).$$

Since

$$\begin{cases} L(0, 0) = K(f) = C \\ \frac{\partial L}{\partial s}(0, 0) = K_*(W_1) \neq 0, \end{cases}$$

the implicit function theorem shows that there is a C^2 function $r \mapsto s(r)$, from a neighborhood of 0 in \mathbb{R} to a neighborhood of 0 in \mathbb{R} , such that

$$(1) \quad C = L(r, s(r)) = K(f + rW + s(r)W_1) \quad \text{for small } r.$$

Notice that the first part of the equation gives, upon differentiating with respect to r ,

$$0 = \frac{\partial L}{\partial r}(0, 0) + \frac{\partial L}{\partial s}(0, 0)s'(0) = K_*(W) + K_*(W_1)s'(0) = K_*(W_1)s'(0),$$

and hence

$$s'(0) = 0.$$

Thus, if we define the variation α by

$$\alpha(u, t) = f(t) + uW(t) + s(u)W_1(t),$$

then each $\bar{\alpha}(u) = f + uW + s(u)W_1$ is in $K^{-1}(C)$ by (1), and also

$$\frac{\partial \alpha}{\partial u}(0, t) = W(t) + s'(0)W_1(t) = W(t). \quad \spadesuit$$

14. **THEOREM (EULER'S RULE).** If the maximum or minimum of J on $K^{-1}(C)$ occurs at a C^2 function f which is not a critical point of K , then there is a number λ such that f is a critical point of $J - \lambda K$ (and consequently the Euler equations for $J - \lambda K$ hold for f).

PROOF. Consider the two linear functions $J_*, K_*: \mathcal{V} \rightarrow \mathbb{R}$. If $K_*(W) = 0$, let α be the variation given by Proposition 13, with all $\bar{\alpha}(u)$ in $K^{-1}(C)$. Since the maximum or minimum of J on $K^{-1}(C)$ occurs at f , the function $u \mapsto J(\bar{\alpha}(u))$ has a maximum or minimum at 0, and consequently

$$J_*(W) = \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} = 0.$$

Thus $\ker K_* \subset \ker J_*$. The vector space \mathcal{V} is infinite dimensional, but it still follows (Problem 3-2) that there is a number λ with $J_* = \lambda K_*$, which is equivalent to the assertion that f is a critical point of $J - \lambda K$. ♦

In Problem I.9-19, we showed that the Euler equations actually make sense and hold for a critical function of J which is only known to be C^1 . A similar result holds for Euler's Rule; because this strengthened form of the rule will be so important for us, the details of the proof will be given here.

Let f be a C^1 function on $[a, b]$, and let W be a C^1 function with $W(a) = W(b) = 0$. Since we no longer have equation (**) on pg. I.319, we can no longer define $J_{f*}(W)$ quite as before. Instead we define

$$J_*(W) = J_{f*}(W) = \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0},$$

where α is the particular variation

$$\alpha(u, t) = f(t) + uW(t).$$

The formula in Problem I.9-19 shows that

$$J_*(W) = \int_a^b W'(t) \left[\frac{\partial F}{\partial y'}(t, f(t), f'(t)) - \int_a^t \frac{\partial F}{\partial x}(t, f(t), f'(t)) dt \right] dt.$$

We define $K_*(W)$ similarly. It is clear that J_* and K_* are linear. Notice that if f is a critical point of K , in the sense that $dK(\bar{\alpha}(u))/du|_{u=0} = 0$ for all variations α keeping endpoints fixed, then surely $K_*(W) = 0$ for all W . Conversely, suppose that $K_*(W) = 0$ for all W . Then Du Bois Reymond's Lemma (see Problem I.9-19) shows that

$$\frac{\partial G}{\partial y'}(t, f(t), f'(t)) - \int_a^t \frac{\partial G}{\partial x}(t, f(t), f'(t)) dt = c,$$

for some constant c . So for any variation α keeping endpoints fixed we have (pg. I.355)

$$\begin{aligned} \left. \frac{dK(\bar{\alpha}(u))}{du} \right|_{u=0} &= c \int_a^b \frac{\partial^2 \alpha}{\partial u \partial t}(0, t) dt = c \left[\frac{\partial \alpha}{\partial u}(0, b) - \frac{\partial \alpha}{\partial u}(0, a) \right] \\ &= 0 - 0. \end{aligned}$$

Thus f is a critical point for K if and only if $K_{f*}(W) = 0$ for all W .

13'. LEMMA. If $K(f) = C$, where the C^1 function f is not a critical point of K , and $K_*(W) = 0$, then $W = \partial \alpha / \partial u(0, t)$ for some variation α [not of the special sort considered above] with the property that each $\bar{\alpha}(u)$ is in $K^{-1}(C)$.

PROOF. The proof of Proposition 13 goes through unchanged; all variations constructed in the proof are of the special sort considered above, except for the final variation α . ♦

14'. THEOREM (EULER'S RULE FOR C^1 FUNCTIONS). If the maximum or minimum of J on $K^{-1}(C)$ occurs at a C^1 function f which is not a critical point of K , then there is a number λ such that f is a critical point of $J - \lambda K$ (and consequently the Euler equations for $J - \lambda K$ make sense and hold for f , by Problem I.9-19).

PROOF. Let \mathcal{W} be the vector space of all C^1 functions W on $[a, b]$ with $W(a) = W(b) = 0$, and consider the two linear functions $J_*, K_*: \mathcal{W} \rightarrow \mathbb{R}$. If $K_*(W) = 0$, let α be the variation given by Proposition 13'. Then the function $u \mapsto J(\bar{\alpha}(u))$ has a maximum or minimum at 0, and consequently

$$\begin{aligned} 0 &= \left. \frac{dJ(\bar{\alpha}(u))}{du} \right|_{u=0} \\ &= \int_a^b W'(t) \left[\frac{\partial F}{\partial y}(t, f(t), f'(t)) - \int_a^t \frac{\partial F}{\partial x}(t, f(t), f'(t)) dt \right] dt \\ &\quad \text{by Problem I.9-19} \\ &= J_*(W). \end{aligned}$$

Thus $\ker K_* \subset \ker J_*$. So there is a number λ with $J_* = \lambda K_*$ on \mathcal{W} . This means that

$$\begin{aligned} &\int_a^b W'(t) \left[\frac{\partial F}{\partial y}(t, f(t), f'(t)) - \int_a^t \frac{\partial F}{\partial x}(t, f(t), f'(t)) dt \right] dt \\ &= \int_a^b W'(t) \lambda \left[\frac{\partial G}{\partial y}(t, f(t), f'(t)) - \int_a^t \frac{\partial G}{\partial x}(t, f(t), f'(t)) dt \right] dt, \end{aligned}$$

for all $W \in \mathcal{W}$. Du Bois Reymond's Lemma then implies, as in the argument preceding Lemma 13', that f is a critical point of $J - \lambda K$. ♦

Notice that, as in the simpler case of functions on \mathbb{R}^n , the dual problem has exactly the same critical points as the original.

Given a certain amount of trust, that similar results hold for functions $f: [a, b] \rightarrow \mathbb{R}^m$, we can finally tackle the classical isoperimetric problem. Consider an imbedding $f: S^1 \rightarrow \mathbb{R}^2$, and let $\alpha: (-\varepsilon, \varepsilon) \times S^1 \rightarrow \mathbb{R}^2$ be a variation of f through imbeddings. For the length $L(\bar{\alpha}(u))$ of $\bar{\alpha}(u)(S^1)$ we have, by formula (I) on page 293,

$$\begin{aligned} \left. \frac{dL(\bar{\alpha}(u))}{du} \right|_{u=0} &= - \int_{S^1} \langle W, \eta \rangle ds \\ &= - \int_{S^1} \langle W, \mathbf{n} \rangle \cdot \kappa ds, \end{aligned}$$

where \mathbf{n} is the principal normal of f and κ is the curvature of f . For the area $A(\bar{\alpha}(u))$ bounded by $\bar{\alpha}(u)(S^1)$ we easily derive, from formula (II) on page 293,

$$\left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_{S^1} \langle W, \mathbf{n} \rangle ds.$$

We want to find the imbedding $f: S^1 \rightarrow \mathbb{R}^2$ which maximizes A for fixed L . Since $f(S^1)$ cannot lie on a straight line, f is not a critical point for L . Therefore Euler's Rule shows that there is some λ with

$$\begin{aligned} 0 &= \int_{S^1} \langle W, \mathbf{n} \rangle ds + \lambda \int_{S^1} \langle W, \mathbf{n} \rangle \cdot \kappa ds \\ &= \int_{S^1} \langle W, \mathbf{n} \rangle [1 + \lambda \kappa] ds, \end{aligned}$$

for all variations W . It clearly follows that we must have $1 + \lambda \kappa = 0$, so κ must be a constant, $-1/\lambda$, and f must be an imbedding as a circle.

[It is perhaps worth pointing out that for this problem one can give an elementary proof that if $L_*(W) = 0$ for some W , then there is a variation $\alpha: (-\varepsilon, \varepsilon) \times S^1 \rightarrow \mathbb{R}^2$ of f with $\partial\alpha/\partial u(0, t) = W$, for which each $\bar{\alpha}(u)(S^1)$ has length $L(0)$. In fact, if β is any variation with $\partial\beta/\partial u(0, t) = W(t)$, then we can set

$$\alpha(u, t) = \frac{L(0)}{L(u)} \cdot \beta(u, t) \in \mathbb{R}^2. \quad L(u) = \text{length of } \bar{\beta}(u)(S^1).$$

We have

$$\begin{aligned}\frac{\partial \alpha}{\partial u}(0, t) &= 1 \cdot \frac{\partial \beta}{\partial u}(0, t) - \frac{L'(0)}{L(0)^2} \cdot \beta(0, t) \\ &= \frac{\partial \beta}{\partial u}(0, t), \quad \text{since } L'(0) = L_*(W) = 0;\end{aligned}$$

and

$$\begin{aligned}\text{length of } \bar{\alpha}(u)(S^1) &= \frac{L(0)}{L(u)} \cdot \text{length of } \bar{\beta}(u)(S^1) \\ &= \frac{L(0)}{L(u)} \cdot L(u) = L(0),\end{aligned}$$

as desired.]

We can also apply Euler's Rule to the dual problem of finding the imbedding $f: S^1 \rightarrow \mathbb{R}^2$ which minimizes L for fixed A . Since no f can be a critical point for A , we find once again that $f(S^1)$ must be a circle. Finally, consider the problem of Dido, to join two fixed points P and Q by a curve c of fixed length $L > d(P, Q)$ so that the area enclosed by c and the line segment \overline{PQ} is a maximum. We consider an imbedding $f: [a, b] \rightarrow \mathbb{R}^2$ with $f(a) = P$ and $f(b) = Q$, and let $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow \mathbb{R}^2$ be a variation of f through imbeddings. For the length $L(\bar{\alpha}(u))$ of $\bar{\alpha}(u)([a, b])$ we have, by formula (I) on page 293,

$$\left. \frac{dL(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_a^b \langle W, \mathbf{n} \rangle \cdot \kappa \, ds,$$

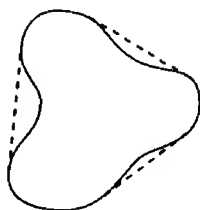
while for the area $A(\bar{\alpha}(u))$ bounded by $\bar{\alpha}(u)([a, b])$ and \overline{PQ} , formula (II) on page 293 reduces to

$$\left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_a^b \langle W, \mathbf{n} \rangle \, ds.$$

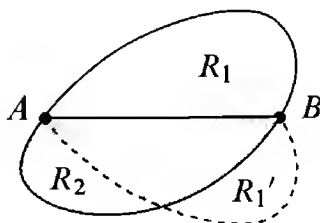
Euler's Rule shows, once again, that f must have constant curvature, so that $f([a, b])$ must be an arc of a circle. We find the same result for the dual problem.

There are, unfortunately, two difficulties with our solution of the isoperimetric problem. We have been working with C^1 curves, and we could have obtained similar results for piecewise C^1 curves with a little more effort. But the obvious class of curves to consider for the isoperimetric problem is the class of rectifiable curves, the curves with finite length (defined as the least upper bound of the lengths of inscribed polygonal curves). Moreover, we have merely found that the circle is the solution of the isoperimetric problem *if a solution exists*; we have not proved that the circle actually is a solution.

Although this will lead us astray from the righteous path of differential geometry, at this point I cannot resist the impulse to mention one of the extremely clever solutions of the isoperimetric problem, involving no assumptions about differentiability, which was given by the great geometer Steiner. Note first that we might as well restrict our attention to convex curves, because the convex hull C^* of a nonconvex curve C has smaller length and encloses a larger area—a suitable region C^{**} similar to C^* will then have the same length as C , and yet still larger area.



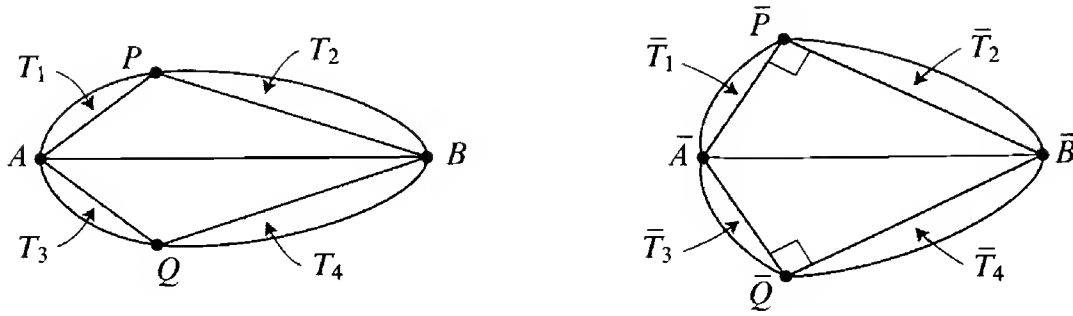
Let us therefore consider a convex curve C which is not a circle. We will show that it cannot be a solution to the isoperimetric problem. Choose two points A and B on C which divide C into two curves C_1 and C_2 of equal length, and let R_i be the region bounded by C_i and the line segment AB . We can assume that $\text{area } R_1 \geq \text{area } R_2$; we claim that we actually have $\text{area } R_1 = \text{area } R_2$. To prove



this, we reflect region R_1 in the line AB , obtaining a region R_1' on the opposite side. Then $R_1 \cup R_1'$ has area \geq the area of $R_1 \cup R_2$, while its circumference is the same. If C is a solution to the isoperimetric problem, then we must actually have $\text{area } R_1 \cup R_1' = \text{area } R_1 \cup R_2$, so we have $\text{area } R_1 = \text{area } R_1' = \text{area } R_2$.

Now since C is not a circle, we can choose A and B so that neither C_1 nor C_2 is a semi-circle. Since $\text{area } R_1 = \text{area } R_2$, the region $R_1 \cup R_1'$ with boundary $C_1 \cup C_1'$ will be another solution to the isoperimetric problem, and it will also not be a circle. In other words, we can assume that C is symmetric with respect to AB .

Now there is a point P on C_1 such that $\angle APB$ is *not* a right angle; let Q be the symmetric point on C_2 . The region inside C is made up of the quadrilateral $APBQ$ together with 4 regions T_1, \dots, T_4 as shown in the left half of the figure on the top of the next page. In the right half of this figure we have drawn



a quadrilateral $\bar{A}\bar{P}\bar{B}\bar{Q}$ with $AP = AQ = \bar{A}\bar{P} = \bar{A}\bar{Q}$ and $BP = BQ = \bar{B}\bar{P} = \bar{B}\bar{Q}$, but with $\angle\bar{A}\bar{P}\bar{B}$ and $\angle\bar{A}\bar{Q}\bar{B}$ both right angles. Then on $\bar{A}\bar{P}$ we have drawn a region \bar{T}_1 congruent to the region T_1 in part (a), and regions $\bar{T}_2, \dots, \bar{T}_4$ have been drawn similarly. The new figure clearly has the same circumference as the original curve C . On the other hand, it has *larger* area, since the quadrilateral $\bar{A}\bar{P}\bar{B}\bar{Q}$ clearly has larger area than $APBQ$. Thus C could not be a solution to the isoperimetric problem. This completes the proof that a circle is the only curve which can be a solution to the isoperimetric problem.

This ingenious proof, although it assumes absolutely nothing about the differentiability of C , still has a defect, which, to be sure, Steiner would never have worried about. This proof, like our previous one, shows only that the circle is the solution of the isoperimetric problem, *if a solution exists*. In Blaschke {1}, {2}, one can find many rigorous solutions of the isoperimetric problem which avoid this pitfall by showing that for a closed curve of length L , enclosing a region of area A , we always have $L^2 - 4\pi A \geq 0$, with equality only when the curve is a circle. These proofs exhibit various degrees of ingenuity and elegance, but there is also a straightforward, if somewhat lengthy, direct proof of existence, which will be useful for us to examine.

Let (X, d) be a bounded metric space, and let $\mathcal{C}(X)$ be the set of all non-empty closed subsets of X . The distance $d(x, C)$ from a point $x \in X$ to a closed set $C \in \mathcal{C}(X)$ is defined as

$$d(x, C) = \min_{y \in C} d(x, y),$$

and we define the ε -neighborhood $V_\varepsilon(C)$ of C as

$$V_\varepsilon(C) = \{x : d(x, C) < \varepsilon\}.$$

Given $C_1, C_2 \in \mathcal{C}(X)$, we then define

$$\rho(C_1, C_2) = \inf \{\varepsilon > 0 : C_1 \subset V_\varepsilon(C_2) \text{ and } C_2 \subset V_\varepsilon(C_1)\}.$$

It is easy to check that ρ is a metric, the **Hausdorff metric**, on $\mathcal{C}(X)$. When X is compact, the corresponding topology on $\mathcal{C}(X)$ depends only on the topology of X , not on the given metric d , since any neighborhood of $C \in \mathcal{C}(X)$ contains an ε -neighborhood.

15. PROPOSITION. If (X, d) is compact, then so is $(\mathcal{C}(X), \rho)$.

PROOF. Given $\varepsilon > 0$, choose a finite number of sets A_1, \dots, A_n of diameter $< \varepsilon$ which cover X . For each finite set $F \subset \{1, \dots, n\}$ let

$$\mathcal{A}_F = \{C \in \mathcal{C}(X) : C \cap A_j \neq \emptyset \iff j \in F\}.$$

Then the sets \mathcal{A}_F cover $\mathcal{C}(X)$ and have diameter $\leq 2\varepsilon$. This shows that $(\mathcal{C}(X), \rho)$ is totally bounded.

Now let C_1, C_2, \dots be a Cauchy sequence in $(\mathcal{C}(X), \rho)$. Let C be the set of all $x \in X$ such that every neighborhood of x contains points from infinitely many C_n . The set C is non-empty, for if $x_n \in C_n$ and x is an accumulation point of the sequence $\{x_n\}$, then $x \in C$. It is also clear that C is closed. We claim that $C = \lim C_n$. Given $\varepsilon > 0$, we first show that the C_n are eventually in the open ε -neighborhood $V_\varepsilon(C)$ of C . For suppose that an infinite sequence C_{i_1}, C_{i_2}, \dots intersected the compact set $X - V_\varepsilon(C)$. Then we could choose $x_{i_n} \in C_{i_n} \cap [X - V_\varepsilon(C)]$; some point $x \in X - V_\varepsilon(C)$ would be an accumulation point of the sequence $\{x_{i_n}\}$, hence $x \in C$, a contradiction. We also claim that C is in $V_\varepsilon(C_n)$ for sufficiently large n . In fact, since C_n is a Cauchy sequence, there is some N such that $C_m \subset V_{\varepsilon/2}(C_n)$ for all $m, n > N$; this implies that $V_{\varepsilon/2}(C_m) \subset V_\varepsilon(C_n)$ for all $m, n > N$. So if $n > N$ and C is not contained in $V_\varepsilon(C_n)$, then also C contains a point which is not in $V_{\varepsilon/2}(C_m)$ for all $m > N$, which is clearly impossible. ♦

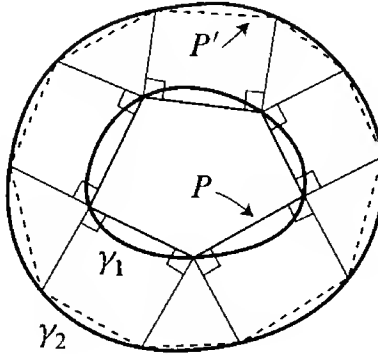
We will apply this result to the case where X is a closed disc in \mathbb{R}^2 . The set $\text{Con}(X) \subset \mathcal{C}(X)$ consisting of all non-empty closed *convex* subsets of X is easily seen to be a closed, and hence compact, subset of $\mathcal{C}(X)$. If $A: \mathcal{C}(X) \rightarrow \mathbb{R}$ is the function $A(C) = \text{area of } C$ ($=$ Lebesgue measure of A , say), then A is clearly continuous. Define $L: \text{Con}(X) \rightarrow \mathbb{R}$ by $L(C) = \text{length of boundary } C$.

16. PROPOSITION. The function $L: \text{Con}(X) \rightarrow \mathbb{R}$ is continuous.

PROOF. If $\rho(C_1, C_2) < \varepsilon$, then $C_1 \subset (1 + \varepsilon) \cdot C_2$ and $C_2 \subset (1 + \varepsilon) \cdot C_1$, so the result follows from

17. LEMMA. If γ_1 and γ_2 are convex curves with γ_1 contained inside γ_2 , then $\text{length } \gamma_1 \leq \text{length } \gamma_2$.

PROOF. The following picture shows that if P is a polygonal arc inscribed

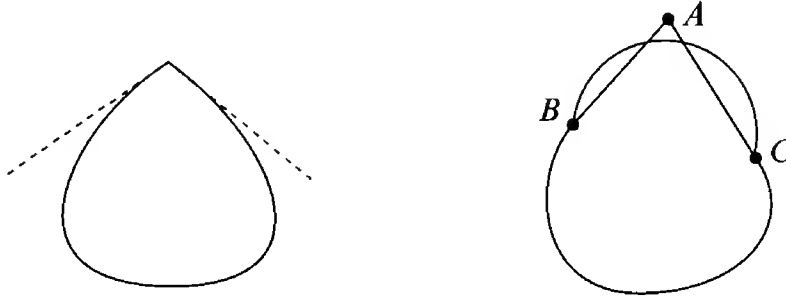


in γ_1 , then P is shorter than some polygonal arc P' inscribed in γ_2 . ♦

It is now an easy matter to prove the existence of a (convex) curve, with fixed length L_0 , of maximum area: We can clearly restrict our attention to convex sets contained within a closed disc X of radius L_0 ; then the set $L^{-1}(L_0) \subset \text{Con}(X)$ is a closed subset of the compact space $\text{Con}(X)$, so the continuous function A takes on its maximum somewhere on the set. This proof of existence, together with Steiner's argument, rigorously solves the isoperimetric problem. The dual problem can be handled similarly. Its solution is also contained in our solution of the original problem, for we now know that the relation $L^2 - 4\pi A \geq 0$ always holds, with equality only for circles, and this proves that the circle is also the solution of the dual problem. It is also easy to derive this fact from the solution of the original problem by using the similarities of the plane. Finally we mention that the problem of Dido can be settled by similar methods; for instance, we can consider the space of all closed convex sets which have a given line segment PQ as part of their boundary:

I would now like to discuss briefly a line of argument which could be used if Steiner's argument were not available, and we had to rely solely on Euler's Rule. Clearly the only problem is to show that the solution of the isoperimetric problem (whose existence we can prove) must be a C^1 curve. The first step would be to show that the solution curve has a tangent line everywhere. Now it is well-known (Problem 2) that a convex function always has left- and right-hand

derivatives, so we just have to show that our convex curve has no corners. If our curve actually contained two straight line segments AB and AC meeting at an angle at A , then it would be easy to show that it is not a solution to the



isoperimetric problem. For the two segments could be replaced by an arc of a circle with equal length, but enclosing larger area, since such an arc is a solution to the problem of Dido. One doesn't really need the whole solution to the problem of Dido to reach this conclusion, however, for a simple calculation will show that the appropriate arc together with line BC encloses more area than triangle ABC . (If we had worked out the calculus of variations argument for piecewise C^1 curves we would have another way of seeing that the two segments can be replaced by some nearby curve of the same length, but enclosing larger area.) In the general case, the same idea can be made to work by an approximation argument.

Now it is also easy to see (Problem 2) that if a convex function is everywhere differentiable, then its derivative is *automatically* continuous. This shows that the solution to the isoperimetric problem must be a C^1 curve; Euler's Rule then leads to the conclusion that it must be a circle.

As differential geometers, we naturally think of generalizing the isoperimetric problem to an arbitrary surface M . Given a variation $\alpha: (-\varepsilon, \varepsilon) \times S^1 \rightarrow M$ of a map $f: S^1 \rightarrow M$ we now have

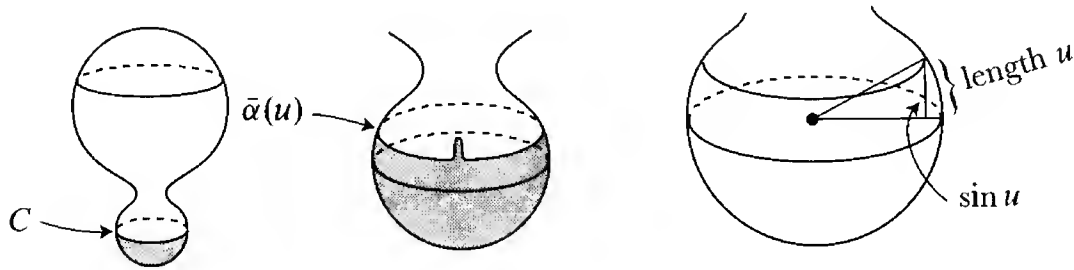
$$\begin{aligned} \left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} &= \int_{S^1} \langle W, \mathbf{u} \rangle ds, \\ \left. \frac{dL(\bar{\alpha}(u))}{du} \right|_{u=0} &= - \int_{S^1} \langle W, \eta \rangle ds \\ &= - \int_{S^1} \langle W, \mathbf{u} \rangle \kappa_g ds, \end{aligned}$$

where \mathbf{u} is the second member of the Darboux frame for f , and κ_g is the geodesic curvature of f . These formulas, together with a few ruthlessly suppressed details which are necessary to transfer Euler's Rule from \mathbb{R}^m to manifolds, show that if f maximizes A for fixed L , then there is a constant λ such that

$$\begin{aligned} 0 &= \int_{S^1} \langle W, \mathbf{u} \rangle ds + \lambda \int_{S^1} \langle W, \mathbf{u} \rangle \kappa_g ds \\ &= \int_{S^1} \langle W, \mathbf{u} \rangle [1 + \lambda \kappa_g] ds \end{aligned}$$

for all variations W . This implies that f has *constant geodesic curvature*. The geodesic curvature was first invented by Minding, in 1830, when he obtained this solution (for the problem of Dido, rather than the isoperimetric problem). Minding dealt with surfaces in \mathbb{R}^3 , and defined κ_g extrinsically, but he then showed that it was a bending invariant; its present name was given it by Bonnet, in 1848.

A rigorous discussion of the isoperimetric problem on an arbitrary surface M is considerably more complicated than for the plane, if for no other reason than because the problem itself is more involved. First of all, Euler's Rule is not always applicable, because there might be closed curves which are geodesics, and consequently critical points for L . For example, on the surface M shown below, the equator C of the smaller spherical part is *not* a critical point for

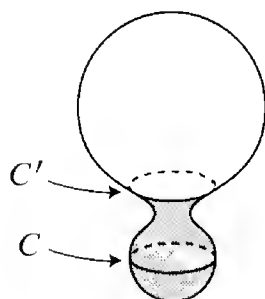


area among all curves with length equal to $L(C)$. We can obtain a variation α of C by moving C up distance u along geodesics perpendicular to C , and then adding on a bulge to bring the length up to $L(C)$. Then $A(u) - A(0)$ is greater than the area of M enclosed between two parallel planes at distance $\sin u$ (see the third picture above), so

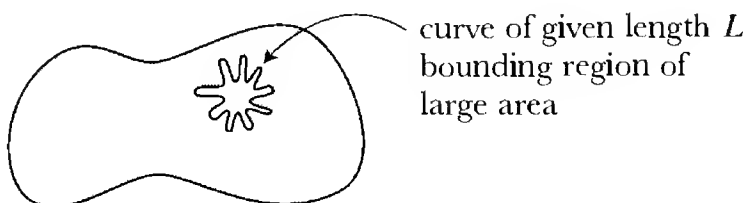
$$A(u) - A(0) > \sin u$$

and consequently $A'(0) \neq 0$.

The figure below shows a curve C' on M , with $L(C') = L(C)$, which is a critical point for A among curves with this length. If we accept the fact that

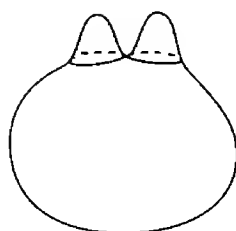


a circle is a solution to the isoperimetric problem on the sphere, then C' must be a solution to the isoperimetric problem on M . Of course, we really have to decide which of the two regions of M bounded by C' should be maximized; if we take the top region, then C' actually minimizes. The necessity of making this decision correctly is further illustrated by the fact that there is another curve C'' higher up with length L that is also a critical point for A among curves of this length. In fact, if we make the wrong decision we might be led to say that there are curves of length L bounding regions with area arbitrarily close to that of M . This becomes quite critical if there is a closed curve of length L which



divides M into two pieces with the same area, as may happen for example on a sphere.

I also suspect that in some cases the solution of the isoperimetric problem will have to be a curve which intersects itself, as in the following picture; notice that

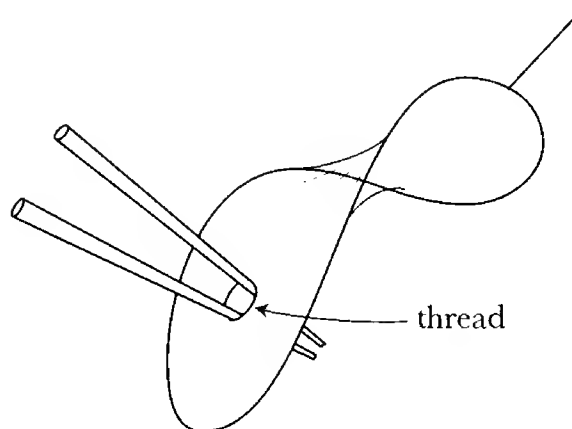


we want the curve to go around as much as possible of the part of the surface with large curvature.

Finally, we point out that on most surfaces there will be just a few solutions of the isoperimetric problem, and that they may be completely different curves. In this respect the problem of Dido is more natural on a general surface; given a geodesic segment γ from P to Q , we would expect that among all curves c from P to Q with given length $L > d(P, Q)$ there is just one on each side of γ which maximizes the area enclosed by γ and c .

I think that a reasonable approach to the isoperimetric problem on a compact surface M is to consider only lengths L so small that a closed curve of length L must be contained in a geodesically convex set. It is then clear that our solution must be the boundary of a geodesically convex set, and there is no problem deciding which region it bounds. All our previous considerations can be suitably modified to show that a solution of the isoperimetric problem exists and is C^1 , so that it must have constant geodesic curvature. This proves, in particular, that there are *closed* curves of constant geodesic curvature; proving this result directly seems almost hopeless. By the way, it is a classical theorem that if *every* curve of constant geodesic curvature is closed, then M must have constant curvature (Blaschke {1}).

In this connection, an interesting experiment can be performed with a soap film on a wire loop. If a small loop of thread held between two thin sticks is dipped into the soap solution, it can then be thrust into the soap film without breaking it. If the part of the soap film inside the thread is then broken, and



the sticks are removed, the thread should take a form which is a solution to the isoperimetric problem on the soap film. When one tries this experiment it turns out that, unless the wire loop is very flat, the string always rushes off toward the wire loop, no matter where it is placed. I take this to mean that there are no curves of constant geodesic curvature on a non-flat minimal surface,

but I haven't the slightest idea how one would prove it. [Actually, as Osserman pointed out to me, the experiment involves a rather more complicated problem, since the shape of the surface *changes* as the thread moves.]

For our next application of Euler's Rule we will work only in \mathbb{R}^3 , and consider the 3-dimensional isoperimetric problem, to find the surface of fixed area A which encloses the greatest volume. Consider an imbedding $f: S^2 \rightarrow \mathbb{R}^3$, and let $\alpha: (-\varepsilon, \varepsilon) \times S^2 \rightarrow \mathbb{R}^3$ be a variation of f through imbeddings. For the area $A(\bar{\alpha}(u))$ of $\bar{\alpha}(u)(S^2)$ we have, by formula (I) on page 294,

$$\begin{aligned} \left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} &= - \int_{S^2} \langle W, \eta \rangle dA \\ &= - \int_{S^2} 2H \langle W, \nu \rangle dA, \end{aligned}$$

where ν is the normal of $f(S^2)$ and H is the mean curvature. For the volume $V(\bar{\alpha}(u))$ enclosed by $\bar{\alpha}(u)(S^2)$ we have, by formula (II) on page 294,

$$\left. \frac{dV(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_{S^2} \langle W, \nu \rangle dA.$$

We want to find the imbedding $f: S^2 \rightarrow \mathbb{R}^3$ which maximizes V for fixed A . Since the compact surface $f(S^2)$ cannot have $H = 0$ everywhere (Corollary 7-31), f is not a critical point for A . Therefore Euler's Rule shows that there is some λ with

$$0 = \int_{S^2} \langle W, \nu \rangle dA + \lambda \int_{S^2} 2H \langle W, \nu \rangle dA = \int_{S^2} \langle W, \nu \rangle [1 + 2\lambda H] dA$$

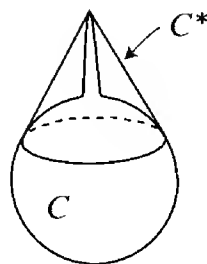
for all variations W . It clearly follows that $f(S^2)$ must have *constant mean curvature*.

At this point we encounter new difficulties, for we first have to find all the surfaces of constant mean curvature. This particular problem is interesting of itself, quite apart from any connection with the isoperimetric problem. For one thing, such surfaces are the possible shape for soap bubbles—the increased air pressure within the bubble naturally makes it take a form which maximizes the enclosed volume. We already know (Theorem 5-3) that a *convex* surface with constant mean curvature must be a standard sphere. H. Hopf [1] proved that an immersed surface homeomorphic to S^2 with constant mean curvature must be a standard sphere. The proof of this is deferred to Addendum 2, since it uses the existence of isothermal parameters, which is proved in Addendum 1. Alexandrov [1] proved that any *imbedded* compact hypersurface of \mathbb{R}^{n+1} with

constant mean curvature must be a standard sphere; this proof is presented in Addendum 3. Alexandrov's theorem holds just as well for hypersurfaces in the hyperbolic space H^{n+1} or in an open hemisphere of S^{n+1} . It definitely fails even for surfaces with $H = 0$ in the sphere S^3 , as we mentioned on page 294.

It was long unknown (Hopf's Problem) whether every *immersed* compact hypersurface with constant mean curvature is a standard sphere, and although this was widely suspected to be the case, the previous edition of this volume mischievously suggested that "some one may some day blow a soap bubble in the shape of an immersed torus". As far as I know, no one has yet done that, but in 1986 Wente [1] proved that there are indeed immersed tori with constant mean curvature. His detailed proof combined methods from complex analysis and recent results on partial differential equations. Noting symmetries in computer-generated pictures of such immersed tori, Abresch [1] searched for examples with one family of planar curvature lines, and was able to reduce the problem to an ODE that can be solved explicitly in terms of elliptic functions. Finally, Kapouleas [1], [2] proved that other surfaces could also be immersed with constant mean curvature.

At first sight the isoperimetric problem seems easier, since it seems that a solution ought to be convex. Proving this directly seems almost hopeless, however, for the boundary of the convex hull C^* of a set C in \mathbb{R}^3 may well have *larger* surface area than the boundary of C . Of course, the volume of C^* is also



larger than that of C —the big question is whether it is larger by enough. In Blaschke {2} there is a proof that the sphere is the solution to the isoperimetric problem provided that we restrict our attention to convex sets. In the general case there is such an overwhelming multitude of problems, not least of which is the difficulty of *defining* surface area, that we will say no more about the problem, merely referring the interested reader to the bibliography.

To conclude this rather disconnected series of remarks, we shall very briefly discuss a problem which requires for its solution even more elaborate machinery than any yet mentioned, but which is of much greater interest to differential

geometry. In his investigations of the “three body problem”, Poincaré was led to consider simple closed geodesics on a compact convex surface $M \subset \mathbb{R}^3$. Poincaré gave a rather long proof that at least one simple closed geodesic exists on M , and then outlined a much more direct argument for the same conclusion. Although many (probably hopelessly difficult) subsidiary results would be required to make this argument into a complete proof, it is nevertheless an interesting application of Euler’s rule for isoperimetric problems. We notice first that if c is a simple closed geodesic on M , then Theorem 6-5 implies that $v \circ c$ divides S^2 into two regions each of area 2π . To establish the existence of such a geodesic, we will consider the set of all simple closed curves γ on M such that $v \circ \gamma$ divides S^2 into two regions of equal area, and then among these choose one, $c: S^1 \rightarrow M$, of shortest length. We claim that c must be a geodesic. To prove this we consider a variation $\alpha: (-\varepsilon, \varepsilon) \times S^1 \rightarrow M$ of c . For the length $L(\bar{\alpha}(u))$ of $\bar{\alpha}(u)(S^1)$ we have, by the formula on page 307,

$$(1) \quad \left. \frac{dL(\bar{\alpha}(u))}{du} \right|_{u=0} = - \int_{S^1} \langle W, \mathbf{u} \rangle \cdot \kappa_g ds.$$

Now extend f to a map $f: D \rightarrow M$, of the unit disc into M , so that $f(D)$ is one of the regions bounded by $f(S^1)$; extend α to a map $\alpha: (-\varepsilon, \varepsilon) \times D \rightarrow M$ similarly. Let $A(\bar{\alpha}(u))$ be the area of the image $v(\bar{\alpha}(u)(D)) \subset S^2$. Then

$$A(\bar{\alpha}(u)) = \int_{\bar{\alpha}(u)(D)} K dA,$$

where dA is the volume element of M and K is the Gaussian curvature of M . It certainly seems reasonable that we should have

$$(2) \quad \left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = \int_{S^1} (K \circ f) \langle W, \mathbf{u} \rangle ds,$$

for $A(\bar{\alpha}(h)) - A(\bar{\alpha}(0))$ is the integral of $K dA$ over a small band around $f(S^1)$ whose width is given approximately by the function $\langle W, \mathbf{u} \rangle$. To prove this rigorously, we write $A(\bar{\alpha}(u))$ as

$$A(\bar{\alpha}(u)) = \int_D [K \circ \bar{\alpha}(u)] \cdot \Gamma(u), \quad \Gamma(u) = \bar{\alpha}(u)^* dA.$$

Then

$$\left. \frac{dA(\bar{\alpha}(u))}{du} \right|_{u=0} = \left. \frac{d}{du} \right|_{u=0} \int_D [K \circ \bar{\alpha}(u)] \cdot \Gamma(u)$$

$$\begin{aligned}
 &= \int_D \left[\frac{d}{du} \Big|_{u=0} K \circ \bar{\alpha}(u) \right] \cdot \Gamma(0) + \int_D (K \circ f) \cdot \dot{\Gamma}(0) \\
 &\quad \text{by Leibnitz's Rule} \\
 &= \int_D \left[\frac{d}{du} \Big|_{u=0} K \circ \bar{\alpha}(u) \right] \cdot \Gamma(0) - \int_D (K \circ f) \langle W, \mathbf{u} \rangle \Gamma(0) \\
 &\quad + \int_{S^1} (K \circ f) (W \lrcorner \Gamma(0)) \quad \text{by Theorem 11} \\
 &= \int_{S^1} (K \circ f) \cdot (W \lrcorner \Gamma(0)) \\
 &= \int_{S^1} (K \circ f) \cdot \langle W, \mathbf{u} \rangle ds.
 \end{aligned}$$

Now if our curve c is a solution to the isoperimetric problem of minimizing L for fixed $A = 2\pi$, then Euler's rule says that there is a constant λ such that

$$0 = \int_{S^1} \langle W, \mathbf{u} \rangle [\lambda(K \circ f) - \kappa_g] ds$$

for all variations W . This implies that $\kappa_g = \lambda(K \circ f)$. On the other hand, applying Theorem 6-5 to $f(D) \subset M$, we obtain

$$- \int_{S^1} \kappa_g ds + 2\pi = \int_{f(D)} K dA = 2\pi,$$

and thus

$$0 = \int_{S^1} \kappa_g ds = \lambda \int_{S^1} (K \circ f) ds.$$

So if M has $K > 0$ everywhere, then we must have $\lambda = 0$, and thus $\kappa_g = 0$; consequently, c is a geodesic.

In Blaschke {1; pp. 211–212} there is a further argument, due to Herglotz, to show that M actually contains at least 3 closed geodesics. Nowadays, all such results are proved by quite different, rigorous methods, of far greater generality—see Klingenberg {1}.

ADDENDUM 1

ISOTHERMAL COORDINATES

As we mentioned in Volume II, the existence of isothermal coordinates on any surface was first proved by Gauss, who resorted to a trick that works only in the analytic case. Although we will treat the more general case also, Gauss' proof will be given first, as it is interesting in its own right. First we need to review some facts about differential equations. The equation

$$y'(x) = f(x, y(x))$$

is written classically as

$$\frac{dy}{dx} = f(x, y),$$

or even as

$$dy - f(x, y) dx = 0.$$

Most elementary differential equations courses indicate that one method of solving this equation is to find an “integrating factor” for it, that is, a nowhere zero function λ such that $\lambda(dy - f dx)$ is exact, say

$$\lambda(dy - f dx) = dg.$$

Then the solutions of the original equation are the same as the solutions of $dg = 0$, i.e., the curves $g = \text{constant}$. For example, to solve the equation

$$\begin{aligned} 0 &= (x^2 y + x) dy + (x y^2 - y) dx \\ &= x dy - y dx + xy(x dy + y dx), \end{aligned}$$

we multiply by $1/xy$, to obtain

$$0 = \frac{dy}{y} - \frac{dx}{x} + d(xy),$$

with the solution

$$\log y(x) - \log x + x \cdot y(x) = \text{constant}.$$

As a more interesting example, we consider the general first order linear equation

$$\frac{dy}{dx} + \phi(x)y = \psi(x),$$

which we write as

$$[\phi(x)y - \psi(x)] dx + dy = 0.$$

In order for

$$[\lambda(x)\phi(x)y - \lambda(x)\psi(x)] dx + \lambda(x) dy$$

to be exact, we need

$$\frac{\partial}{\partial y} [\lambda(x)\phi(x)y - \lambda(x)\psi(x)] = \frac{d\lambda}{dx}$$

or

$$\begin{aligned} \lambda(x)\phi(x) &= \frac{d\lambda}{dx} \implies \frac{d\lambda}{\lambda} = \phi(x) dx \\ &\implies \log \lambda = \int \phi \\ &\implies \lambda = e^{\int \phi}. \end{aligned}$$

So we write our original equation as

$$e^{\int \phi} \frac{dy}{dx} + e^{\int \phi} \phi(x)y = \psi(x)e^{\int \phi},$$

which gives

$$\begin{aligned} \frac{d}{dx} (e^{\int \phi} y) &= \psi(x)e^{\int \phi}, \\ e^{\int \phi} y &= \int e^{\int \phi} \psi + C, \\ y &= e^{-\int \phi} \left(\int e^{\int \phi} \psi + C \right). \end{aligned}$$

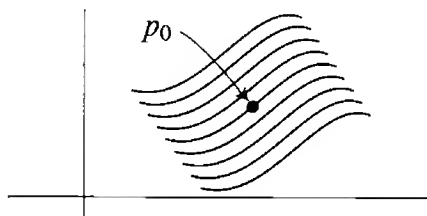
Of course, only in the most fortuitous cases can one find an integrating factor by inspection. What is theoretically more interesting is the observation that for any 1-form

$$(*) \quad \omega = \alpha dx + \beta dy$$

on \mathbb{R}^2 with $\alpha(p_0), \beta(p_0) \neq 0$, an integrating factor exists in a neighborhood of p_0 . To prove this, we consider the differential equation

$$(**) \quad y'(x) = -\frac{\beta}{\alpha}(x, y(x)).$$

Since $-(\beta/\alpha)(p_0) \neq 0$, the integral curves of this differential equation form a foliation in a neighborhood of p_0 and there is a diffeomorphism h from a



neighborhood of p_0 to \mathbb{R}^2 such that the integral curves go into the sets with 2nd coordinate constant. Let

$$g(p) = 2^{\text{nd}} \text{ coordinate of } h(p).$$

Then

$$\begin{aligned} \ker dg(p) &= \text{tangent space at } p \text{ of the solution curve} \\ &\quad \text{of } (**) \text{ going through } p \\ &= \ker \omega(p). \end{aligned}$$

This proves that

$$dg(p) = \lambda(p) \cdot \omega(p)$$

for some $\lambda(p) \neq 0$.

In Problem I.6-9 we showed that the differential equation

$$y'(z) = f(z, y(z)) \quad (' = \text{complex derivative})$$

can always be solved if $f: \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}$ is complex analytic. From this we easily conclude, by modifying the preceding argument, that if α and β are complex-valued functions on \mathbb{R}^2 which are the restrictions of complex analytic functions on \mathbb{C}^2 , and $\alpha(p_0), \beta(p_0) \neq 0$, then there is a complex-valued function λ in a neighborhood of p_0 such that

$$\lambda(\alpha dx + \beta dy) = dg$$

for some complex-valued function g ; both λ and g are the restrictions of complex analytic functions on \mathbb{C}^2 . Now we can prove

18. **THEOREM.** Let $\langle \cdot, \cdot \rangle$ be a Riemannian metric on a neighborhood V of $0 \in \mathbb{R}^2$ whose components g_{ij} with respect to the standard coordinate system on \mathbb{R}^2 are C^ω (= real analytic). Then there exists a C^ω isothermal coordinate system for $\langle \cdot, \cdot \rangle$ in a neighborhood of 0 .

PROOF. Let X_1, X_2 be a C^ω orthonormal moving frame in a neighborhood of 0 , with dual 1-forms θ^1, θ^2 . Then

$$\langle \cdot, \cdot \rangle = \theta^1 \otimes \theta^1 + \theta^2 \otimes \theta^2,$$

and consequently the corresponding quadratic function $\| \cdot \|^2$ can be written as

$$\| \cdot \|^2 = \theta^1 \cdot \theta^1 + \theta^2 \cdot \theta^2.$$

Let ϕ be the complex-valued differential form

$$\phi = \theta^1 + i\theta^2, \quad \text{with} \quad \bar{\phi} = \theta^1 - i\theta^2.$$

Then

$$\| \cdot \|^2 = \phi \cdot \bar{\phi}.$$

If we construct X_1, X_2 explicitly by applying the Gram-Schmidt orthonormalization process to $\partial/\partial x^1, \partial/\partial x^2$, then the coefficients of X_1, X_2 will appear as algebraic combinations of the g_{ij} . The same is thus true of θ^1, θ^2 . Since the g_{ij} are C^ω , and hence the restrictions of complex analytic functions on \mathbb{C}^2 , the same is true for θ^1, θ^2 . So by the remark preceding the theorem, there is a complex-valued function λ such that

$$\lambda \phi = dg \implies \bar{\lambda} \bar{\phi} = d\bar{g}$$

for some complex-valued function g . This implies that

$$\lambda \bar{\lambda} \| \cdot \|^2 = \lambda \bar{\lambda} \phi \cdot \bar{\phi} = dg \cdot d\bar{g},$$

so that

$$\| \cdot \|^2 = \frac{1}{\lambda \bar{\lambda}} dg \cdot d\bar{g}.$$

If we write $g = u + iv$ for real-valued u and v , then the Jacobian of $(u, v): \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is not zero, for if it were, then dg would be zero, and hence $\| \cdot \|^2$ would be zero. Now

$$dg \cdot d\bar{g} = du \cdot du + dv \cdot dv,$$

so

$$\| \cdot \|^2 = \frac{1}{\lambda \bar{\lambda}} (du \cdot du + dv \cdot dv).$$

By polarization,

$$\langle \cdot, \cdot \rangle = \frac{1}{\lambda \bar{\lambda}} (du \otimes du + dv \otimes dv).$$

The functions u and v are C^ω since g is the restriction of a complex analytic function on \mathbb{C}^2 . Thus (u, v) is the required C^ω isothermal coordinate system. ♦

The proof of Theorem 18 when the g_{ij} are not C^ω will be **much** more involved. First we introduce some new classes of functions. A function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to satisfy a **Hölder condition of order α** ($0 < \alpha < 1$) on $U \subset \mathbb{R}^n$ if there is a constant K such that

$$|f(p) - f(q)| \leq K \cdot |p - q|^\alpha \quad \text{for all } p, q \in U.$$

Such functions are called C^α functions, and a function f is $C^{n+\alpha}$ if all mixed n^{th} order derivatives of f exist and are C^α . We will eventually show that if the g_{ij} in Theorem 18 are C^α , then there is a $C^{1+\alpha}$ isothermal coordinate system in a neighborhood of 0. We will also show that if the g_{ij} are $C^{n+\alpha}$, then this same coordinate system is $C^{n+1+\alpha}$; in particular, if the g_{ij} are C^∞ , so is the coordinate system. There need not be a C^1 isothermal coordinate system when the g_{ij} are merely C^0 (= continuous).

The condition that (u, v) be isothermal is

$$\sum_{i,j} g_{ij} dx^i \otimes dx^j = \langle \ , \ \rangle = \lambda(du \otimes du + dv \otimes dv), \quad \text{some } \lambda > 0.$$

To derive explicit equations for u and v , it is easiest to consider the dual metric $\langle \ , \ \rangle^*$ on $T^*\mathbb{R}^2$, which must satisfy

$$\begin{aligned} \sum_{i,j} g^{ij} \left(\frac{\partial}{\partial x^i} \right)^{**} \otimes \left(\frac{\partial}{\partial x^j} \right)^{**} &= \langle \ , \ \rangle^* \\ &= \frac{1}{\lambda} \left[\left(\frac{\partial}{\partial u} \right)^{**} \otimes \left(\frac{\partial}{\partial u} \right)^{**} + \left(\frac{\partial}{\partial v} \right)^{**} \otimes \left(\frac{\partial}{\partial v} \right)^{**} \right]. \end{aligned}$$

Denoting (x^1, x^2) by (x, y) , setting

$$(g^{ij}) = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

and applying our equation to the pairs (du, du) , (dv, dv) , and (du, dv) , we obtain

$$(1) \quad au_x^2 + 2bu_xu_y + cu_y^2 = \frac{1}{\lambda} = av_x^2 + 2bv_xv_y + cv_y^2$$

$$(2) \quad au_xv_x + b(u_xv_y + u_yv_x) + cu_yv_y = 0.$$

Equation (2) can be written

$$u_x(av_x + bv_y) + u_y(bv_x + cv_y) = 0,$$

which implies that there is a function ρ with

$$\begin{aligned} u_x &= \rho(bv_x + cv_y) \\ u_y &= -\rho(av_x + bv_y). \end{aligned}$$

Substituting into (1) we find that

$$\rho^2(ac - b^2) = 1.$$

We thus have the

Beltrami equations:

$$u_x = \frac{bv_x + cv_y}{\sqrt{ac - b^2}}, \quad u_y = -\frac{av_x + bv_y}{\sqrt{ac - b^2}}, \quad \begin{pmatrix} a & b \\ b & c \end{pmatrix} = (g^{ij})$$

as necessary and sufficient conditions that (u, v) be isothermal coordinates for the metric $\langle \cdot, \cdot \rangle = \sum_{i,j=1}^n g_{ij} dx^i \otimes dx^j$.

At this point it becomes extremely convenient to introduce the notation of formal complex derivatives. We will often denote a typical point of $\mathbb{C} = \mathbb{R}^2$ by z , and z will also be used to denote the identity map $z: \mathbb{C} \rightarrow \mathbb{C}$. We have already used $x, y: \mathbb{R}^2 \rightarrow \mathbb{R}$ for the coordinate functions on \mathbb{R}^2 , so the equation $z = x + iy$ is a (true) equation concerning the three complex-valued functions x, y, z on \mathbb{R}^2 . Because of this equation we have

$$dz = dx + i dy, \quad d\bar{z} = dx - i dy.$$

Since any complex-valued differential on \mathbb{R}^2 can be written in terms of dx and dy , it can also be written in terms of dz and $d\bar{z}$. So for any complex-valued function w on \mathbb{R}^2 , there are unique functions $w_z = \partial w / \partial z$ and $w_{\bar{z}} = \partial w / \partial \bar{z}$ with

$$dw = w_z dz + w_{\bar{z}} d\bar{z}.$$

Substituting from the above equations, we have

$$dw = (w_z + w_{\bar{z}}) dx + i(w_z - w_{\bar{z}}) dy,$$

so that

$$\begin{aligned} w_z + w_{\bar{z}} &= \frac{\partial w}{\partial x} = w_x \\ i(w_z - w_{\bar{z}}) &= \frac{\partial w}{\partial y} = w_y, \end{aligned}$$

which gives

$$\boxed{\begin{aligned} w_z &= \frac{1}{2}(w_x - iw_y) \\ w_{\bar{z}} &= \frac{1}{2}(w_x + iw_y) \end{aligned}} \quad \text{or} \quad \boxed{\begin{aligned} w_x &= w_z + w_{\bar{z}} \\ w_y &= \frac{w_{\bar{z}} - w_z}{i}. \end{aligned}}$$

The usual differentiation rules apply to the operators $\partial/\partial z$ and $\partial/\partial \bar{z}$, and we have

$$\begin{aligned} \frac{\partial}{\partial z}(z) &= 1, & \frac{\partial}{\partial \bar{z}}(z) &= 0 \\ \frac{\partial}{\partial \bar{z}}(\bar{z}) &= 1, & \frac{\partial}{\partial z}(\bar{z}) &= 0. \end{aligned}$$

It is also easy to check that we always have

$$w_{z\bar{z}} = w_{\bar{z}z}.$$

Another easily checked result is

$$\bar{w}_{\bar{z}} = \overline{(w_z)}.$$

The chain rule becomes

$$\begin{aligned} (w \circ \zeta)_z &= (w_z \circ \zeta) \cdot \zeta_z + (w_{\bar{z}} \circ \zeta) \cdot \bar{\zeta}_z \\ (w \circ \zeta)_{\bar{z}} &= (w_z \circ \zeta) \cdot \zeta_{\bar{z}} + (w_{\bar{z}} \circ \zeta) \cdot \bar{\zeta}_{\bar{z}}. \end{aligned}$$

[If we agree to write $w_z \circ \zeta = w_\zeta$ and $w_{\bar{z}} \circ \zeta = w_{\bar{\zeta}}$, then we have

$$\begin{aligned} (w \circ \zeta)_z &= w_\zeta \cdot \zeta_z + w_{\bar{\zeta}} \cdot \bar{\zeta}_z \\ (w \circ \zeta)_{\bar{z}} &= w_\zeta \cdot \zeta_{\bar{z}} + w_{\bar{\zeta}} \cdot \bar{\zeta}_{\bar{z}}, \end{aligned}$$

which looks a little nicer.]

Finally, we note that if $w = u + iv$ for real-valued u and v , then the condition

$$0 = w_{\bar{z}} = \frac{1}{2}(u_x + iv_x + i[u_y + iv_y])$$

is equivalent to the Cauchy-Riemann equations

$$u_x = v_y, \quad u_y = -v_x.$$

So $w_{\bar{z}} = 0$ if and only if w is complex analytic on U ; in this case it is also easy to see that

$$w_z = w', \quad \text{the complex derivative.}$$

Now suppose that u, v satisfy the Beltrami equations. If we set

$$w = u + iv,$$

we find that

$$\begin{aligned} 2w_{\bar{z}}\sqrt{ac - b^2} &= (b - ia + i\sqrt{ac - b^2})v_x + (c - ib - \sqrt{ac - b^2})v_y, \\ 2w_z\sqrt{ac - b^2} &= (b + ia + i\sqrt{ac - b^2})v_x + (c + ib + \sqrt{ac - b^2})v_y. \end{aligned}$$

A short calculation shows that the coefficients of v_x and v_y on the right hand sides of these two equations are proportional, and we have

$$\frac{w_{\bar{z}}}{w_z} = \frac{c - a - 2ib}{c + a + 2\sqrt{ac - b^2}}$$

or

$$(*) \quad w_{\bar{z}} = \mu w_z, \quad \mu = \frac{c - a - 2ib}{c + a + 2\sqrt{ac - b^2}}.$$

Conversely, it is easy to see that the Beltrami equations follow from (*). Notice that if the g_{ij} are $C^{n+\alpha}$, then so are a, b, c and hence μ . Moreover, $|\mu| < 1$. Notice also that we always have

$$u_x v_y - u_y v_x = |w_z|^2 - |w_{\bar{z}}|^2.$$

So if w satisfies (*), then

$$u_x v_y - u_y v_x = |w_z|^2(1 - |\mu|^2).$$

Since $|\mu| < 1$, it follows that (u, v) has non-zero Jacobian at any point where $w_z \neq 0$.

The first major step on the road to our final result will be to prove that if μ is C^α and $|\mu(0)| < 1$, then equation (*) has a $C^{1+\alpha}$ solution w in a neighborhood of 0, with $w_z(0) \neq 0$. In outline our proof will go as follows. We will let $D(R)$

denote the open disc of radius $R > 0$. Suppose that f is C^α in $D(R)$. For all $z_0 \in D(R)$, define

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z)}{z - z_0} dx dy \quad (z = x + iy).$$

We will show that

$$(A) \quad F_{\bar{z}}(z_0) = f(z_0).$$

We thus have a way of producing a function F with $F_{\bar{z}} = f$.

Now suppose for the moment that we have a function w satisfying (*). If we define

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z - z_0} dx dy \quad z_0 \in D(R),$$

then (A) gives

$$F_{\bar{z}}(z_0) = \mu(z_0)w_z(z_0) = w_{\bar{z}}(z_0).$$

But this means that $(w - F)_{\bar{z}} = 0$, so $w - F$ is complex analytic. Thus we have

$$w(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z - z_0} dx dy + g(z_0),$$

for some complex analytic function g . Conversely, if w satisfies this integral equation for some complex analytic function g , then (A) shows that w satisfies (*), since $g_{\bar{z}} = 0$. We will solve (*) by showing that the equivalent integral equation always has a solution.

In order to get to the proof of (A), we need a succession of simple lemmas.

19. LEMMA (GENERALIZED CAUCHY INTEGRAL THEOREM). Let $D \subset \mathbb{R}^2$ be a compact 2-dimensional manifold-with-boundary, and let $f: D \rightarrow \mathbb{C}$ be C^1 . Then

$$\int_{\partial D} f dz = 2i \iint_D f_{\bar{z}} dx dy.$$

PROOF. If $f = u + iv$ for C^1 functions $u, v: D \rightarrow \mathbb{R}$, then

$$\int_{\partial D} f dz = \int_{\partial D} (u + iv)(dx + i dy) = \int_{\partial D} u dx - v dy + i \int_{\partial D} v dx + u dy,$$

while

$$\begin{aligned} 2i \iint_D f_{\bar{z}} dx dy &= 2i \iint_D \frac{1}{2}(f_x + i f_y) dx dy = \iint_D (-f_y + i f_x) dx dy \\ &= \iint_D (-u_y - v_x) dx dy + i \iint_D (u_x - v_y) dx dy. \end{aligned}$$

The real and imaginary parts of these two expressions are equal by Stokes' Theorem. ♦

Remark: We define the line integral $\int_c f d\bar{z}$ as

$$\int_c f d\bar{z} = \int_c f \cdot (dx - i dy).$$

It is easy to check that this definition is equivalent to the one usually adopted in complex analysis books,

$$\int_c f d\bar{z} = \overline{\left(\int_c \bar{f} dz \right)}.$$

Since

$$\bar{f}_{\bar{z}} = \overline{(f_z)},$$

Lemma 19 gives

$$\begin{aligned} \int_{\partial D} f d\bar{z} &= \overline{\left(\int_{\partial D} \bar{f} dz \right)} = -2i \overline{\left(\iint_D \bar{f}_{\bar{z}} dx dy \right)} \\ &= -2i \iint_D f_z dx dy. \end{aligned}$$

20. LEMMA (GENERALIZED CAUCHY INTEGRAL FORMULA). For f and D as in Lemma 19, and $z_0 \in \text{interior } D$, we have

$$f(z_0) = \frac{1}{2\pi i} \int_{\partial D} \frac{f(z)}{z - z_0} dz - \frac{1}{\pi} \iint_D \frac{f_{\bar{z}}(z)}{z - z_0} dx dy.$$

PROOF. Let $B(\varepsilon) \subset D$ be a disc of radius ε around z_0 . Applying Lemma 19 to the function

$$z \mapsto \frac{f(z)}{z - z_0} \quad \text{on} \quad D - \text{interior } B(\varepsilon),$$

we have

$$2i \iint_{D - \text{int } B(\varepsilon)} \frac{f_{\bar{z}}(z)}{z - z_0} dx dy = \int_{\partial D} \frac{f(z)}{z - z_0} dz + \int_{\partial B(\varepsilon)} \frac{f(z)}{z - z_0} dz.$$

Taking the limit as $\varepsilon \rightarrow 0$, we find that

$$2i \iint_D \frac{f_{\bar{z}}(z)}{z - z_0} dx dy = \int_{\partial D} \frac{f(z)}{z - z_0} dz + 2\pi i f(z_0). \quad \spadesuit$$

21. LEMMA. If $z_0 \in D(R)$, then

$$\bar{z}_0 = -\frac{1}{\pi} \iint_{D(R)} \frac{1}{z - z_0} dx dy.$$

PROOF. Let $\overline{D(R)}$ be the closure of $D(R) \subset \mathbb{C}$. Applying Lemma 20 to $\bar{z}: \overline{D(R)} \rightarrow \mathbb{C}$ we have

$$\bar{z}_0 = \frac{1}{2\pi i} \int_{\partial \overline{D(R)}} \frac{\bar{z}}{z - z_0} dz - \frac{1}{\pi} \iint_{D(R)} \frac{1}{z - z_0} dx dy,$$

and

$$\int_{\partial \overline{D(R)}} \frac{\bar{z}}{z - z_0} dz = \int_{\partial \overline{D(R)}} \frac{R^2}{z(z - z_0)} dz = 0,$$

since the sum of the residues of $R^2/z(z - z_0)$ inside $\partial \overline{D(R)}$ is 0. \spadesuit

22. LEMMA. If $z_0 \in D(R)$, then

$$|z_0|^2 = -\frac{1}{\pi} \iint_{D(R)} \frac{z}{z - z_0} dx dy + R^2.$$

PROOF. Since $|z|^2 = z\bar{z}$, so that

$$\frac{\partial |z|^2}{\partial \bar{z}} = z,$$

Lemma 20 now gives

$$|z_0|^2 = \frac{R^2}{2\pi i} \int_{\partial D(R)} \frac{1}{z - z_0} dz - \frac{1}{\pi} \int_{D(R)} \frac{z}{z - z_0} dx dy,$$

and

$$\frac{R^2}{2\pi i} \int_{\partial D(R)} \frac{1}{z - z_0} dz = R^2. \diamond$$

And now one somewhat more technical lemma.

23. LEMMA. Let $0 < \varepsilon_1, \varepsilon_2 \leq 1$, with $\varepsilon_1 + \varepsilon_2 \neq 2$. Then there is a constant $c(\varepsilon_1, \varepsilon_2)$, not depending on R , such that

$$\iint_{D(R)} \frac{dx dy}{|z - z_1|^{2-\varepsilon_1} \cdot |z - z_2|^{2-\varepsilon_2}} \leq c(\varepsilon_1, \varepsilon_2) \cdot \frac{1}{|z_1 - z_2|^{2-\varepsilon_1-\varepsilon_2}}$$

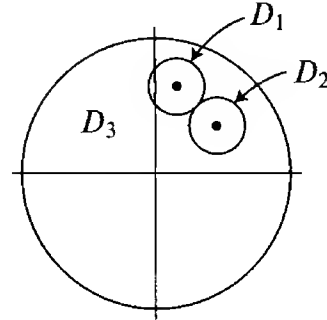
for all $z_1, z_2 \in D(R)$ with $z_1 \neq z_2$.

PROOF. Let $|z_1 - z_2| = 2\delta$ and define

D_1 = disc of radius δ around z_1

D_2 = disc of radius δ around z_2

$D_3 = D(R) - (D_1 \cup D_2)$.



Clearly

$$\begin{aligned} \iint_{D_1} \frac{dx dy}{|z - z_1|^{2-\varepsilon_1} \cdot |z - z_2|^{2-\varepsilon_2}} &\leq \frac{1}{\delta^{2-\varepsilon_2}} \iint_{D_1} \frac{dx dy}{|z - z_1|^{2-\varepsilon_1}} \\ &= \frac{1}{\delta^{2-\varepsilon_2}} \int_0^{2\pi} \int_0^\delta \frac{1}{r^{2-\varepsilon_1}} \cdot r dr d\theta && \text{using polar} \\ &= \frac{1}{\delta^{2-\varepsilon_2}} \cdot 2\pi \cdot \int_0^\delta r^{\varepsilon_1-1} dr && \text{coordinates} \\ &= \frac{1}{\delta^{2-\varepsilon_2}} \cdot 2\pi \cdot \frac{\delta^{\varepsilon_1}}{\varepsilon_1} = \frac{2\pi}{\varepsilon_1} \cdot \frac{1}{\delta^{2-\varepsilon_1-\varepsilon_2}}. \end{aligned}$$

Similarly, the integral over D_2 is

$$\leq \frac{2\pi}{\varepsilon_2} \cdot \frac{1}{\delta^{2-\varepsilon_1-\varepsilon_2}}.$$

These bounds both have the desired form

$$c(\varepsilon_1, \varepsilon_2) \cdot \frac{1}{|z_1 - z_2|^{2-\varepsilon_1-\varepsilon_2}}.$$

Now we always have

$$|z - z_1| \leq |z - z_2| + |z_1 - z_2| = |z - z_2| + 2\delta,$$

and consequently

$$\frac{|z - z_1|}{|z - z_2|} \leq 1 + \frac{2\delta}{|z - z_2|} \quad z \neq z_2.$$

So on D_3 (in fact on $\mathbb{R}^2 - D_2$) we have

$$\frac{|z - z_1|}{|z - z_2|} \leq 1 + \frac{2\delta}{\delta} = 3.$$

So

$$\begin{aligned} \iint_{D_3} \frac{dx \, dy}{|z - z_1|^{2-\varepsilon_1} \cdot |z - z_2|^{2-\varepsilon_2}} &= \iint_{D_3} \left| \frac{z - z_1}{z - z_2} \right|^{2-\varepsilon_2} \cdot \frac{dx \, dy}{|z - z_1|^{4-\varepsilon_1-\varepsilon_2}} \\ &\leq 3^{2-\varepsilon_2} \iint_{D_3} \frac{dx \, dy}{|z - z_1|^{4-\varepsilon_1-\varepsilon_2}} \\ &\leq 3^{2-\varepsilon_2} \iint_{\mathbb{R}^2 - D_1} \frac{dx \, dy}{|z - z_1|^{4-\varepsilon_1-\varepsilon_2}} \\ &\leq 3^{2-\varepsilon_2} \int_0^{2\pi} \int_\delta^\infty r^{\varepsilon_1+\varepsilon_2-3} \, dr \quad \text{using polar} \\ &\quad \text{coordinates} \\ &\quad \text{around } z_1 \\ &= 3^{2-\varepsilon_2} \cdot 2\pi \cdot \frac{1}{2-\varepsilon_1-\varepsilon_2} \cdot \delta^{\varepsilon_1+\varepsilon_2-2}, \end{aligned}$$

which is again of the desired form. ♦

We are now ready to give the precise formulation of (A), which includes three inequalities that are essential for proving that the equivalent integral equation can be solved.

24. PROPOSITION. Let $f: D(R) \rightarrow \mathbb{C}$ satisfy

$$\begin{aligned} |f(z)| &\leq M & z \in D(R) \\ |f(z_1) - f(z_2)| &\leq K|z_1 - z_2|^\alpha & z_1, z_2 \in D(R). \end{aligned}$$

Define

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z)}{z - z_0} dx dy, \quad z_0 \in D(R).$$

Then

$$\begin{aligned} \text{(a)} \quad & F_{\bar{z}}(z_0) = f(z_0) \\ \text{(b)} \quad & F_z(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2} dx dy. \end{aligned}$$

Moreover for all $z_0, z_1, z_2 \in D(R)$ we have

$$\begin{aligned} \text{(c)} \quad & |F(z_0)| \leq 4RM \\ \text{(d)} \quad & |F_z(z_0)| \leq \frac{2^{\alpha+1}}{\alpha} R^\alpha K \\ \text{(e)} \quad & |F_z(z_1) - F_z(z_2)| \leq CK|z_1 - z_2|^\alpha, \end{aligned}$$

where C is a constant that *does not depend on R , or on the function f .*

PROOF. For fixed z_0 , let

$$\tilde{F}(z) = F(z) - f(z_0)\bar{z},$$

so that by Lemma 21

$$\tilde{F}(z') = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{z - z'} dx dy.$$

We claim that the complex derivative $\tilde{F}'(z_0)$ exists and that in fact

$$\tilde{F}'(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2} dx dy.$$

To prove this we have to show that as $h \rightarrow 0$, the same is true of

$$\begin{aligned}
& \left| \frac{\tilde{F}(z_0 + h) - \tilde{F}(z_0)}{h} + \frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2} dx dy \right| \\
&= \left| -\frac{1}{\pi h} \iint_{D(R)} [f(z) - f(z_0)] \cdot \left\{ \frac{1}{z - z_0 - h} - \frac{1}{z - z_0} \right\} dx dy \right. \\
&\quad \left. + \frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2} dx dy \right| \\
&= \left| -\frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0 - h)(z - z_0)} + \frac{1}{\pi} \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2} dx dy \right| \\
&= \frac{1}{\pi} \left| \iint_{D(R)} \frac{f(z) - f(z_0)}{z - z_0} \left\{ \frac{1}{z - z_0 - h} - \frac{1}{z - z_0} \right\} dx dy \right| \\
&= \frac{|h|}{\pi} \left| \iint_{D(R)} \frac{f(z) - f(z_0)}{(z - z_0)^2(z - z_0 - h)} dx dy \right| \\
&\leq \frac{|h|}{\pi} \iint_{D(R)} \frac{|f(z) - f(z_0)|}{|z - z_0|^2 |z - z_0 - h|} dx dy \\
&\leq \frac{K|h|}{\pi} \iint_{D(R)} \frac{|z - z_0|^\alpha}{|z - z_0|^2 |z - z_0 - h|} dx dy \\
&= \frac{K|h|}{\pi} \iint_{D(R)} \frac{dx dy}{|z - z_0|^{2-\alpha} |z - z_0 - h|} \\
&\leq \frac{K}{\pi} |h| \cdot c(\alpha, 1) \cdot |h|^{\alpha+1-2} \quad \text{by Lemma 23} \\
&= \frac{K}{\pi} c(\alpha, 1) \cdot |h|^\alpha.
\end{aligned}$$

This indeed approaches 0 as $h \rightarrow 0$.

Now since the complex derivative $\tilde{F}'(z_0)$ exists for all $z_0 \in D(R)$, the ordinary partials F_x, F_y exist, and hence \tilde{F}_z and $\tilde{F}_{\bar{z}}$ exist. Moreover, since \tilde{F} is complex analytic, from the definition of \tilde{F} we obtain

$$0 = \tilde{F}_{\bar{z}}(z_0) = F_{\bar{z}}(z_0) - f(z_0) \cdot 1,$$

which proves (a). Furthermore

$$\tilde{F}'(z_0) = \tilde{F}_z(z_0) = F_z(z_0) - 0,$$

which proves (b).

To prove (c), we note that

$$\begin{aligned} |F(z_0)| &\leq \frac{M}{\pi} \iint_{D(R)} \frac{1}{|z - z_0|} dx dy \\ &\leq \frac{M}{\pi} \iint_D \frac{1}{|z - z_0|} dx dy && \text{where } D \supset D(R) \text{ is the disc} \\ &= \frac{M}{\pi} \int_0^{2\pi} \int_0^{2R} \frac{1}{r} \cdot r dr d\theta && \text{using polar} \\ &= 4RM. && \text{coordinates} \\ & && \text{around } z_0 \end{aligned}$$

Similarly, for (d) we have

$$\begin{aligned} |F_z(z_0)| &= \left| \frac{1}{\pi} \iint_{D(R)} \frac{|f(z) - f(z_0)|}{|z - z_0|^2} dx dy \right| && \text{by (b)} \\ &\leq \frac{K}{\pi} \iint_{D(R)} \frac{1}{|z - z_0|^{2-\alpha}} dx dy \\ &\leq \frac{K}{\pi} \iint_D \frac{1}{|z - z_0|^{2-\alpha}} dx dy \\ &= \frac{K}{\pi} \int_0^{2\pi} \int_0^{2R} \frac{1}{r^{2-\alpha}} r dr d\theta \\ &= \frac{2K(2R)^\alpha}{\alpha}. \end{aligned}$$

To prove (e) let z_1, z_2 be fixed, and define

$$\begin{cases} B = \frac{f(z_1) - f(z_2)}{z_1 - z_2} \\ \tilde{F}(z) = F(z) - Bz\bar{z}. \end{cases}$$

If we set

$$\tilde{f}(z) = f(z) - Bz,$$

then by Lemma 22 we have

$$\tilde{F}(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\tilde{f}(z)}{z - z_0} dx dy - BR^2.$$

So by (b) we have

$$\tilde{F}_z(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\tilde{f}(z) - \tilde{f}(z_0)}{(z - z_0)^2} dx dy.$$

Thus

$$\tilde{F}_z(z_1) - \tilde{F}_z(z_2) = -\frac{1}{\pi} \iint_{D(R)} \left\{ \frac{\tilde{f}(z) - \tilde{f}(z_1)}{(z - z_1)^2} - \frac{\tilde{f}(z) - \tilde{f}(z_2)}{(z - z_2)^2} \right\} dx dy.$$

But we easily check that

$$\tilde{f}(z_1) = \tilde{f}(z_2).$$

Therefore

$$\begin{aligned} & \tilde{F}_z(z_1) - \tilde{F}_z(z_2) \\ &= -\frac{1}{\pi} \iint_{D(R)} [\tilde{f}(z) - \tilde{f}(z_1)] \cdot \left\{ \frac{1}{(z - z_1)^2} - \frac{1}{(z - z_2)^2} \right\} dx dy \\ &= -\frac{1}{\pi} \iint_{D(R)} \frac{[\tilde{f}(z) - \tilde{f}(z_1)] \cdot (z_1 - z_2)(2z - z_1 - z_2)}{(z - z_1)^2(z - z_2)^2} dx dy \\ &= -\frac{1}{\pi} \iint_{D(R)} \frac{[\tilde{f}(z) - \tilde{f}(z_1)] \cdot (z_1 - z_2)[(z - z_1) + (z - z_2)]}{(z - z_1)^2(z - z_2)^2} dx dy. \end{aligned}$$

Now since

$$\begin{aligned} \tilde{f}(z) - \tilde{f}(z_1) &= f(z) - f(z_1) - B(z - z_1) \\ &\parallel \\ \tilde{f}(z) - \tilde{f}(z_2) &= f(z) - f(z_2) - B(z - z_2), \end{aligned}$$

we have

$$\begin{aligned}
 & [\tilde{f}(z) - \tilde{f}(z_1)][(z - z_1) + (z - z_2)] \\
 &= [f(z) - f(z_1) - B(z - z_1)](z - z_2) \\
 &\quad + [f(z) - f(z_2) - B(z - z_2)](z - z_1) \\
 &= [f(z) - f(z_1)](z - z_2) + [f(z) - f(z_2)](z - z_1) \\
 &\quad - 2B(z - z_1)(z - z_2).
 \end{aligned}$$

So we get

$$\begin{aligned}
 \tilde{F}_z(z_1) - \tilde{F}_z(z_2) &= -\frac{(z_1 - z_2)}{\pi} \iint_{D(R)} \frac{f(z) - f(z_1)}{(z - z_1)^2(z - z_2)} dx dy \\
 &\quad - \frac{(z_1 - z_2)}{\pi} \iint_{D(R)} \frac{f(z) - f(z_2)}{(z - z_1)(z - z_2)^2} dx dy \\
 &\quad + \frac{2B}{\pi} \iint_{D(R)} \frac{(z_1 - z_2)}{(z - z_1)(z - z_2)} dx dy \\
 &= I_1 + I_2 + I_3, \quad \text{say.}
 \end{aligned}$$

Now

$$\begin{aligned}
 |I_1| &\leq \frac{|z_1 - z_2|}{\pi} \iint_{D(R)} \frac{K}{|z - z_1|^{2-\alpha} \cdot |z - z_2|} dx dy \\
 &\leq \frac{|z_1 - z_2|}{\pi} \frac{K \cdot c(\alpha, 1)}{|z_1 - z_2|^{1-\alpha}} \quad \text{by Lemma 23} \\
 &= \frac{K \cdot c(\alpha, 1)}{\pi} |z_1 - z_2|^\alpha.
 \end{aligned}$$

Similarly,

$$|I_2| \leq \frac{K \cdot c(\alpha, 1)}{\pi} |z_1 - z_2|^\alpha.$$

Finally,

$$\begin{aligned}
 I_3 &= \frac{2B}{\pi} \iint_{D(R)} \left(\frac{1}{z - z_1} - \frac{1}{z - z_2} \right) dx dy \\
 &= 2B(\bar{z}_1 - \bar{z}_2) \quad \text{by Lemma 21,}
 \end{aligned}$$

so

$$|I_3| \leq 2|B| \cdot |z_1 - z_2|.$$

We have

$$|B| = \frac{|f(z_1) - f(z_2)|}{|z_1 - z_2|} \leq K|z_1 - z_2|^{\alpha-1}.$$

Therefore

$$|I_3| \leq 2K|z_1 - z_2|^\alpha.$$

Thus

$$\begin{aligned} |\tilde{F}_z(z_1) - \tilde{F}_z(z_2)| &\leq |I_1| + |I_2| + |I_3| \\ &\leq (\text{constant}) \cdot K \cdot |z_1 - z_2|^\alpha. \end{aligned}$$

From the definition of \tilde{F} we have

$$\tilde{F}_z(z) = F_z(z) - B\bar{z},$$

so we have

$$\begin{aligned} |F_z(z_1) - F_z(z_2)| &\leq (\text{constant}) \cdot K \cdot |z_1 - z_2|^\alpha + |B| \cdot |\bar{z}_1 - \bar{z}_2| \\ &\leq (\text{constant}) \cdot K \cdot |z_1 - z_2|^\alpha + K|z_1 - z_2|^{\alpha-1} \cdot |z_1 - z_2|, \\ &\quad \text{by the estimate for } |B| \text{ above} \\ &\leq CK|z_1 - z_2|^\alpha. \quad \spadesuit \end{aligned}$$

Instead of solving the equation

$$(*) \quad w_{\bar{z}} = \mu w_z.$$

or the equivalent integral equation, for reasons that will appear later we will instead solve the more general equation

$$(**) \quad w_{\bar{z}} = \mu w_z + \gamma w + \delta.$$

where μ, γ, δ are C^α and $|\mu(0)| < 1$; moreover, we will show that solutions exist with any given values for $w(0)$ and $w_z(0)$. There is no loss of generality in assuming that $\mu(0) = 0$:

25. LEMMACHEN. If the equation

$$(**) \quad w_{\bar{z}} = \mu w_z + \gamma w + \delta$$

has a $C^{1+\alpha}$ solution in a neighborhood of 0, with arbitrary values for $w(0)$ and $w_z(0)$, for all C^α functions μ, γ, δ with $\mu(0) = 0$, then it also has such $C^{1+\alpha}$ solutions for all C^α functions μ, γ, δ with $|\mu(0)| < 1$.

PROOF. For any function w , define \tilde{w} by

$$\tilde{w}(z) = w(z - \mu(0)\bar{z}),$$

so that

$$w(z) = \tilde{w}(z + \mu(0)\bar{z}).$$

The chain rule on page 320 gives

$$(1) \quad \begin{aligned} w_z(z) &= \tilde{w}_z + \tilde{w}_{\bar{z}} \cdot \overline{\mu(0)}, \\ w_{\bar{z}}(z) &= \tilde{w}_z \cdot \mu(0) + \tilde{w}_{\bar{z}}, \end{aligned}$$

where $\tilde{w}_z, \tilde{w}_{\bar{z}}$ are to be evaluated at $z + \mu(0)\bar{z}$. Therefore

$$(**) \quad w_{\bar{z}} = \mu w_z + \gamma w + \delta$$

if and only if

$$\mu(0) \cdot \tilde{w}_z + \tilde{w}_{\bar{z}} = \mu(z)[\tilde{w}_z + \overline{\mu(0)}\tilde{w}_{\bar{z}}] + \gamma(z)w(z) + \delta(z),$$

or

$$\begin{aligned} \tilde{w}_{\bar{z}} &= \left(\frac{\mu(z) - \mu(0)}{1 - \overline{\mu(0)}\mu(z)} \right) \tilde{w}_z + \frac{\gamma(z)}{1 - \overline{\mu(0)}\mu(z)} w(z) + \frac{\delta(z)}{1 - \overline{\mu(0)}\mu(z)} \\ &= \rho(z)\tilde{w}_{\bar{z}} + \sigma(z)w(z) + \tau(z), \quad \text{say,} \end{aligned}$$

where $\rho(0) = 0$. In this equation $\tilde{w}_{\bar{z}}, \tilde{w}_z$ are evaluated at $z + \overline{\mu(0)}z$. Replacing z by $z - \overline{\mu(0)}z$, we get the equivalent equation

$$(\widetilde{**}) \quad \tilde{w}_{\bar{z}}(z) = \rho(z - \overline{\mu(0)}z)\tilde{w}_{\bar{z}}(z) + \sigma(z - \overline{\mu(0)}z)\tilde{w}(z) + \tau(z - \overline{\mu(0)}z),$$

which is of the same form as $(**)$, with the coefficient of $\tilde{w}_{\bar{z}}$ being 0 at 0. So by hypothesis we can solve for a $C^{1+\alpha}$ function \tilde{w} with any desired initial values

$$\tilde{w}(0) = \tilde{a}, \quad \tilde{w}_z(0) = \tilde{b}.$$

This gives

$$w(0) = \tilde{w}(0) = \tilde{a},$$

while by equation (1)

$$w_z(0) = \tilde{w}_z(0) + \overline{\mu(0)}\tilde{w}_{\bar{z}}(0).$$

Using equation (**), we have

$$\tilde{w}_{\bar{z}}(0) = \sigma(0)\tilde{w}(0) + \tau(0) = \sigma(0) \cdot \tilde{a} + \tau(0),$$

so

$$w_z(0) = \tilde{b} + \overline{\mu(0)}[\sigma(0) \cdot \tilde{a} + \tau(0)].$$

So to solve (**) for

$$w(0) = a, \quad w_z(0) = b,$$

we just solve (**) for

$$\begin{aligned} \tilde{w}(0) &= a \\ \tilde{w}_z(0) &= b - \overline{\mu(0)}[\sigma(0)\tilde{a} + \tau(0)]. \quad \spadesuit \end{aligned}$$

Since we will be solving the general equation

$$(**) \quad w_{\bar{z}} = \mu w_z + \gamma w + \delta, \quad \mu(0) = 0,$$

we first want to find an integral equation equivalent to it. To do this we form

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z) + \gamma(z)w(z) + \delta(z)}{z - z_0} dx dy.$$

Proposition 24 gives

$$F_{\bar{z}} = \mu w_z + \gamma w + \delta = w_{\bar{z}} \quad \text{if } w \text{ satisfies } (**),$$

and hence $(w - F)_{\bar{z}} = 0$, so that

$$\begin{aligned} w(z_0) &= -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z - z_0} dx dy - \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z - z_0} dx dy \\ &\quad - \frac{1}{\pi} \iint_{D(R)} \frac{\delta(z)}{z - z_0} dx dy + g(z_0) \end{aligned}$$

for some complex analytic function g . Conversely, of course, if w satisfies this equation for a complex analytic g , then it satisfies (**). By complicating our integral equation, we can arrange that $w(0) = g(0)$; clearly we just have to add

$$\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z} dx dy + \cdots$$

to the right hand side. Similarly, if we add

$$z \cdot \left\{ \frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z^2} + \cdots \right\}$$

to the right hand side, we will have $w_z(0) = g'(0)$; this follows from Proposition 24 (and the fact that $\mu(0) = 0$). So we see that we can solve (**) for w with any given values of $w(0), w_z(0)$ provided that we can solve the following equation for w , where g is any complex analytic function (actually it would suffice to solve it for functions of the form $g(z) = \tilde{a}z + \tilde{b}$):

$$\begin{aligned} w(z_0) &= -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z - z_0} dx dy - \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z - z_0} dx dy - \frac{1}{\pi} \iint_{D(R)} \frac{\delta(z)}{z - z_0} dx dy \\ &+ \frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\delta(z)}{z} dx dy \\ &+ z_0 \left\{ \frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z^2} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z^2} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\delta(z)}{z^2} dx dy \right\} \\ &+ g(z_0). \end{aligned}$$

Now the first integral involving δ is a $C^{1+\alpha}$ function Δ , for Proposition 24 shows that

$$\begin{aligned} \Delta_{\bar{z}} &= \delta && \text{which is } C^\alpha \text{ by assumption,} \\ \Delta_z &\text{ is } C^\alpha && \text{by part (e) of Proposition 24.} \end{aligned}$$

The other two integrals involving δ are just constants. So it certainly suffices

to show that we can solve the following equation for C^α functions μ, γ with $\mu(0) = 0$ and any $C^{1+\alpha}$ function h :

$$(I) \quad \begin{aligned} w(z_0) = & -\frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z - z_0} dx dy - \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z - z_0} dx dy \\ & + \frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z} dx dy \\ & + z_0 \left\{ \frac{1}{\pi} \iint_{D(R)} \frac{\mu(z)w_z(z)}{z^2} dx dy + \frac{1}{\pi} \iint_{D(R)} \frac{\gamma(z)w(z)}{z^2} dx dy \right\} \\ & + h(z_0). \end{aligned}$$

The integral equation (I) will be solved by the only method available to us, namely, the method of successive approximation, which we have always formulated in terms of the Contraction Lemma (I.5-1). First we need to concoct the right space of functions to work with. Consider first the set

$$H(R, \alpha) = \{C^\alpha \text{ functions } w: D(R) \rightarrow \mathbb{C}\}.$$

For $w \in H(R, \alpha)$ we define

$$\|w\|_R = \sup_{z \in D(R)} |w(z)| + R^\alpha \cdot \sup_{\substack{z_1, z_2 \in D(R) \\ z_1 \neq z_2}} \frac{|w(z_1) - w(z_2)|}{|z_1 - z_2|^\alpha}.$$

The first term $\sup |w(z)|$ insures that $w_n \rightarrow 0$ uniformly if $\|w_n\|_R \rightarrow 0$. The term $\sup |w(z_1) - w(z_2)|/|z_1 - z_2|^\alpha$ is simply the “best” constant K in the definition of w being C^α ; the fudge factor R^α is reasonable, for it insures that

$$\|w\|_R = \|\tilde{w}\|_1 \quad \text{where} \quad \tilde{w}(z) = w(Rz).$$

It is easy to check that

$$\begin{aligned} \|w\|_R &> 0 && \text{for } w \neq 0, \\ \|\lambda w\|_R &= |\lambda| \cdot \|w\|_R && \lambda \in \mathbb{R}, \\ \|w_1 + w_2\|_R &\leq \|w_1\|_R + \|w_2\|_R. \end{aligned}$$

So we obtain a metric on $H(R, \alpha)$ by defining

$$\text{distance from } w_1 \text{ to } w_2 = \|w_1 - w_2\|_R,$$

and it is easy to see that $H(R, \alpha)$ is complete in this metric. Finally, it is easily checked that if $w_1, w_2 \in H(R, \alpha)$, then

$$\|w_1 w_2\|_R \leq \|w_1\|_R \cdot \|w_2\|_R.$$

Next consider the set

$$H(R, \alpha + 1) = \{C^{\alpha+1} \text{ functions } w: D(R) \rightarrow \mathbb{C}\}.$$

For $w \in H(R, \alpha + 1)$ we define

$$\|w\|_R = \sup_{z \in D(R)} |w(z)| + R \cdot \|w_z\|_R + R \cdot \|w_{\bar{z}}\|_R.$$

It is once again easy to check that

$$\|w\|_R > 0 \quad \text{for } w \neq 0$$

$$\|\lambda w\|_R = |\lambda| \cdot \|w\|_R$$

$$\|w_1 + w_2\|_R \leq \|w_1\|_R + \|w_2\|_R,$$

and that $H(R, \alpha + 1)$ is a complete metric space with respect to the metric

$$\text{distance from } w_1 \text{ to } w_2 = \|w_1 - w_2\|_R.$$

Consider, for the moment, a function $f: (-R, R) \rightarrow \mathbb{R}$, and suppose that

$$R^{\alpha+1} \cdot \frac{|f'(x_1) - f'(x_2)|}{|x_1 - x_2|^\alpha} \leq K.$$

Defining

$$g(s) = f(x_1 + s(x_2 - x_1)) - f(x_2 + s(x_1 - x_2)),$$

we have

$$g(0) = f(x_1) - f(x_2), \quad g(1) = f(x_2) - f(x_1).$$

So the mean value theorem gives

$$\begin{aligned} 2[f(x_2) - f(x_1)] &= \frac{g(1) - g(0)}{1 - 0} = g'(\xi) & \xi \in (0, 1) \\ &= (x_2 - x_1) \cdot [f'(\eta_1) - f'(\eta_2)] & \eta_1, \eta_2 \in (x_1, x_2). \end{aligned}$$

Thus

$$\begin{aligned} R^\alpha |f(x_1) - f(x_2)| &\leq \frac{R^\alpha |x_1 - x_2|}{2} \cdot \frac{K |\eta_1 - \eta_2|^\alpha}{R^{\alpha+1}} \\ &= \frac{K}{2} \frac{|x_1 - x_2|}{R} \cdot |\eta_1 - \eta_2|^\alpha \\ &\leq K \cdot |x_1 - x_2|^\alpha, \end{aligned}$$

or finally

$$R^\alpha \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|^\alpha} \leq K.$$

For functions $w: D(R) \rightarrow \mathbb{C}$ there is a similar argument, using Taylor's formula to estimate $|g(1) - g(0)|$. The answer involves the derivative Dw , which can be expressed in terms of w_z and $w_{\bar{z}}$. From this argument we easily see that there is an inequality of the form

$$\|w\|_R \leq (\text{constant}) \cdot |||w|||_R.$$

26. PROPOSITION. Let μ, γ be C^α functions in a neighborhood of 0 with $\mu(0) = 0$, and let h be $C^{\alpha+1}$ in a neighborhood of 0. Then for sufficiently small $R > 0$ there is a $C^{\alpha+1}$ function $w: D(R) \rightarrow \mathbb{C}$ satisfying (I) for all $z_0 \in D(R)$.

PROOF. We suppose that μ, γ are C^α in $D(R_0)$ for some $R_0 \leq 1$, and we will henceforth consider only $R \leq R_0$. For $w \in H(R, \alpha + 1)$, define the function Sw on $D(R)$ by setting $(Sw)(z_0)$ equal to the right side of (I) without the $h(z_0)$: we will abbreviate this expression by

$$\begin{aligned} (Sw)(z_0) = & I_1(z_0) + I_2(z_0) \\ & + I_3(z_0) + I_4(z_0) \\ & + z_0\{I_5(z_0) + I_6(z_0)\}. \end{aligned}$$

We make the crucial

CLAIM. There is a constant C' , depending only on α , and not on R , such that

$$|||Sw|||_R \leq C' \cdot R^\alpha \cdot |||w|||_R$$

for all $w \in H(R, \alpha + 1)$.

Assuming this Claim for the moment, the remainder of the proof goes as follows. Since $R^\alpha \rightarrow 0$ as $R \rightarrow 0$, there is clearly some R_* such that for all $R \leq R_*$ we have

$$|||Sw|||_R \leq C'' \cdot |||w|||_R,$$

where C'' is a constant with

$$C'' < 1, \quad \frac{|||h|||_R}{3}.$$

Define $T : H(R, \alpha + 1) \rightarrow H(R, \alpha + 1)$ by

$$Tw = Sw + h.$$

If $R \leq R_*$, then for all w with

$$\|w\|_R \leq \frac{3}{2} \|h\|_R$$

we have

$$\begin{aligned} \|Tw\|_R &= \|Sw + h\|_R \leq \|Sw\|_R + \|h\|_R \\ &\leq \frac{\|h\|_R}{3} \cdot \|w\|_R + \|h\|_R \\ &\leq \frac{1}{2} \|h\|_R + \|h\|_R \\ &= \frac{3}{2} \|h\|_R. \end{aligned}$$

Thus, for $R \leq R_*$, the map T takes the complete metric space

$$M = \left\{ w \in H(R, \alpha + 1) : \|w\|_R \leq \frac{3}{2} \|h\|_R \right\}$$

into itself. Moreover, the map $T : M \rightarrow M$ is a contraction, for

$$\begin{aligned} \|Tw_1 - Tw_2\|_R &= \|Sw_1 - Sw_2\|_R \\ &= \|S(w_1 - w_2)\|_R \leq C'' \cdot \|w_1 - w_2\|_R. \end{aligned}$$

By the Contraction Lemma, there is some $w \in M$ with

$$w = Tw = Sw + h,$$

which is precisely the equation we want.

To prove the Claim we will use all the information in Lemma 24. First we want to show that

$$\|I_1\|_R \leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R,$$

where the constant is independent of R . It clearly suffices to prove the same inequality for each of

$$\sup |I_1(z)|, \quad R \cdot \|(I_1)_z\|_R, \quad R \cdot \|(I_1)_{\bar{z}}\|_R.$$

Let L be a number such that

$$|\mu(z_1) - \mu(z_2)| \leq L \cdot |z_1 - z_2|^\alpha, \quad z_1, z_2 \in D(R_0).$$

Since $\mu(0) = 0$, it follows that

$$|\mu(z)| \leq L R^\alpha, \quad z \in D(R), \quad R \leq R_0$$

and therefore that

$$\|\mu\|_R \leq 2L R^\alpha.$$

Thus for all $z, z_1, z_2 \in D(R)$ we have

$$\begin{aligned} (1) \quad |\mu(z)w_z(z)| &\leq \|\mu w_z\|_R \leq \|\mu\|_R \cdot \|w_z\|_R \\ &\leq 2L R^\alpha \cdot \frac{\|w\|_R}{R} \\ &= 2L R^{\alpha-1} \cdot \|w\|_R, \\ (2) \quad \frac{|\mu(z_1)w_z(z_1) - \mu(z_2)w_z(z_2)|}{|z_1 - z_2|^\alpha} &\leq \frac{\|\mu w_z\|_R}{R^\alpha} \leq \frac{\|\mu\|_R \cdot \|w_z\|_R}{R^\alpha} \\ &\leq \frac{2L R^\alpha \cdot \|w\|_R}{R^\alpha \cdot R} \\ &= \frac{2L}{R} \cdot \|w\|_R. \end{aligned}$$

We can now apply the inequalities of Proposition 24. Inequality (c) gives

$$\begin{aligned} (3) \quad |I_1(z)| &\leq 4R \cdot 2L R^{\alpha-1} \cdot \|w\|_R \\ &= 8L \cdot R^\alpha \cdot \|w\|_R, \end{aligned}$$

which is the desired inequality for $\sup |I_1(z)|$. Inequalities (d) and (e) give

$$\begin{aligned} (4) \quad R \cdot |(I_1)_z(z)| &\leq R \frac{2^{\alpha+1}}{\alpha} R^\alpha \cdot \frac{2L}{R} \|w\|_R \\ &= \frac{2^{\alpha+2}}{\alpha} L \cdot R^\alpha \cdot \|w\|_R \end{aligned}$$

$$\begin{aligned} (5) \quad R^{\alpha+1} \cdot \frac{|(I_1)_z(z_1) - (I_1)_z(z_2)|}{|z_1 - z_2|^\alpha} &\leq R^{\alpha+1} \cdot C \cdot \frac{2L}{R} \|w\|_R \\ &= 2CL \cdot R^\alpha \cdot \|w\|_R; \end{aligned}$$

these give the desired inequality for $R \cdot \|(I_1)_z\|_R$. Finally, since $(I_1)_{\bar{z}} = \mu z$, the necessary inequalities for $R \cdot \|(I_1)_{\bar{z}}\|_R$ follow immediately from (1), (2). We have

therefore shown that

$$\|I_1\|_R \leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R.$$

Now consider I_2 . We first note that for $z \in D(R)$ we have

$$\begin{aligned} |\gamma(z)w(z)| &\leq \|\gamma w\|_R \leq \|\gamma\|_R \cdot \|w\|_R \\ &\leq \|\gamma\|_{R_0} \cdot (\text{constant}) \cdot \|w\|_R \quad (\text{see page 338}). \end{aligned}$$

This is a *stronger* inequality than (1): since $0 < R \leq 1$ and $0 < \alpha < 1$ we have $1 \leq R^{\alpha-1}$, so we can write

$$(1') \quad |\gamma(z)w(z)| \leq (\text{constant}) \cdot R^{\alpha-1} \cdot \|w\|_R.$$

Similarly, if $z_1, z_2 \in D(R)$, then

$$\begin{aligned} (2') \quad \frac{|\gamma(z_1)w(z_1) - \gamma(z_2)w(z_2)|}{|z_1 - z_2|^\alpha} &\leq \frac{\|\gamma w\|_R}{R^\alpha} \leq \frac{\|\gamma\|_R \cdot \|w\|_R}{R^\alpha} \\ &\leq \frac{\|\gamma\|_{R_0}}{R^\alpha} \cdot (\text{constant}) \cdot \|w\|_R \\ &\leq \frac{\text{constant}}{R} \cdot \|w\|_R. \end{aligned}$$

Now (1'), (2') give the inequality

$$\|I_2\|_R \leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R$$

in the same way that (1), (2) gave the inequality for $\|I_1\|_R$.

Since I_3 is just a constant, $I_3(z) = I_1(0)$, we have

$$\begin{aligned} \|I_3\|_R &= \|I_1(0)\|_R = |I_1(0)| \leq \sup_{z \in D(R)} |I_1(z)| \\ &\leq \|I_1\|_R \leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R. \end{aligned}$$

Similarly for I_4 .

As for the term $zI_5(z) = z(I_1)_z(0)$, we have

$$\begin{aligned} |z(I_1)_z(0)| &\leq |z| \cdot |(I_1)_z(0)| \\ &\leq R \cdot \frac{\|I_1\|_R}{R} = \|I_1\|_R \\ &\leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R, \\ R \cdot \|\{z \cdot (I_1)_z(0)\}_z\|_R &= R \cdot \|(I_1)_z(0)\|_R \\ &= R \cdot |(I_1)_z(0)| \\ &\leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R. \quad \text{as above.} \end{aligned}$$

Thus $\|z(I_1)_z(0)\|_R \leq (\text{constant}) \cdot R^\alpha \cdot \|w\|_R$, and the term $zI_6(z)$ works exactly the same. ♦

27. COROLLARY. If μ, γ, δ are C^α functions in a neighborhood of 0, with $|\mu(0)| < 1$, and $a, b \in \mathbb{C}$ are any two complex numbers, then there is a $C^{\alpha+1}$ function w in a neighborhood of 0 such that

$$\begin{aligned} w_{\bar{z}} &= \mu w_z + \gamma w + \delta \\ (**) \quad w(0) &= a \\ w_z(0) &= b. \end{aligned}$$

In particular, there is a $C^{1+\alpha}$ isothermal coordinate system around any point of a surface with a C^α metric.

PROOF. Proposition 26 and Lemmacheen 25. ♦

Our next task is to show that if μ, γ, δ in Corollary 27 are $C^{n+\alpha}$, then there is a solution w of $(**)$ which is $C^{n+1+\alpha}$. First a technical lemma.

28. LEMMA. If f is $C^{n+\alpha}$ ($n \geq 1$) on $D(R)$ and we define

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{f(z)}{z - z_0} dx dy, \quad z_0 \in D(R),$$

then F is $C^{n+1+\alpha}$.

PROOF. Induction on n . Consider first the case $n = 1$. By Proposition 24 we have $F_{\bar{z}} = f$, so $F_{\bar{z}}$ is $C^{1+\alpha}$. We just have to show that F_z is $C^{1+\alpha}$, since this then implies that F_x, F_y are $C^{1+\alpha}$, so that F is $C^{2+\alpha}$. Now we easily check that

$$\frac{\partial}{\partial z} \log |z - z_0|^2 = \frac{1}{z - z_0},$$

and therefore

$$F(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\partial}{\partial z} (f \log |z - z_0|^2) dx dy + \frac{1}{\pi} \iint_{D(R)} f_z \log |z - z_0|^2 dx dy.$$

Using the Remark after Lemma 19, we write this as

$$F(z_0) = \frac{1}{2\pi i} \int_{\partial \overline{D(R)}} f \log |z - z_0|^2 d\bar{z} + \frac{1}{\pi} \iint_{D(R)} f_z \log |z - z_0|^2 dx dy.$$

We can now differentiate under the integral signs to obtain

$$(1) \quad F_z(z_0) = \frac{1}{2\pi i} \int_{\partial D(R)} \frac{f(z)}{z - z_0} d\bar{z} + \frac{1}{\pi} \iint_{D(R)} \frac{f_z(z)}{z - z_0} dx dy.$$

The first integral is C^∞ (since we can keep differentiating under the integral sign); the second is $C^{1+\alpha}$ by Proposition 24.

Now suppose the result holds for $C^{n+\alpha}$ functions, and let f be $C^{n+1+\alpha}$. We still have $F_{\bar{z}} = f$, so that $F_{\bar{z}}$ is $C^{n+1+\alpha}$, and we also have equation (1), in which the first integral is C^∞ . Now f_z is $C^{n+\alpha}$, so by the induction assumption, the second integral is $C^{n+1+\alpha}$. Thus F_z is $C^{n+1+\alpha}$, so F is $C^{n+2+\alpha}$. ♦

29. PROPOSITION. If μ, γ, δ are $C^{n+\alpha}$ functions in a neighborhood of 0, with $|\mu(0)| < 1$, and $a, b \in \mathbb{C}$ are any two complex numbers, then there is a $C^{n+1+\alpha}$ function w in a neighborhood of 0 such that

$$\begin{aligned} w_{\bar{z}} &= \mu w_z + \gamma w + \delta \\ (**) \quad w(0) &= a \\ w_z(0) &= b. \end{aligned}$$

In particular, there is a $C^{n+1+\alpha}$ isothermal coordinate system around any point of a surface with a $C^{n+\alpha}$ metric.

PROOF. Induction on n . The case $n = 0$ is Corollary 26. Now suppose the result is true for n , and let μ, γ, δ be $C^{n+1+\alpha}$.

Case 1. $\gamma = 0$. The motivation for the proof is the following. If w satisfies

$$(1) \quad w_{\bar{z}} = \mu w_z + \delta,$$

then we should have

$$(w_z)_{\bar{z}} = w_{\bar{z}z} = \mu(w_z)_z + \mu_z w_z + \delta_z.$$

So we first solve this equation for w_z . To be precise, we note that μ, μ_z, δ_z are $C^{n+\alpha}$, so since the result is assumed true for n , there is a function f satisfying

$$(2) \quad f_{\bar{z}} = \mu f_z + \mu_z f + \delta_z$$

in some disc $D(R)$; moreover, we can obtain any desired values for $f(0)$ and $f_z(0)$. [Notice that equation (2) contains f explicitly even though equation (1) does not contain w explicitly.] Define W by

$$\bar{W}(z_0) = -\frac{1}{\pi} \iint_{D(R)} \frac{\bar{f}(z)}{z - z_0} dx dy.$$

Then W is $C^{n+2+\alpha}$ by Lemma 28, and by Proposition 24 we have

$$\bar{f}(z_0) = \bar{W}_{\bar{z}}(z_0) = \overline{W_z(z_0)} \implies f(z_0) = W_z(z_0).$$

So

$$\begin{aligned} (W_{\bar{z}})_z &= W_{z\bar{z}} = \bar{f}_z = \mu f_z + \mu_z f + \delta_z \quad \text{by (1)} \\ &= (\mu f)_z + \delta_z = (\mu W_z)_z + \delta_z. \end{aligned}$$

Hence $(W_{\bar{z}} - \mu W_z - \delta)_z = 0$. This means that we can write

$$(3) \quad W_{\bar{z}}(z) - \mu(z)W_z(z) - \delta(z) = g(\bar{z}),$$

where g is complex analytic. Let G be a complex analytic function with $G_{\bar{z}}(\bar{z}) = g(\bar{z})$, and let

$$w(z) = W(z) - G(\bar{z}).$$

Then

$$\begin{aligned} w_z &= W_z - 0 \\ w_{\bar{z}}(z) &= W_{\bar{z}}(z) - g(\bar{z}) = \mu(z)W_z(z) + \delta(z) \quad \text{by (3)} \\ &= \mu(z)w_z(z) + \delta(z). \end{aligned}$$

Thus w is a $C^{n+2+\alpha}$ solution of our equation. We also have

$$\begin{aligned} w(0) &= W(0) - G(0) \\ w_z(0) &= W_z(0) = f(0). \end{aligned}$$

So we obtain the condition $w_z(0) = b$ by choosing a solution f of (2) with $f(0) = b$. We can obtain $w(0) = a$ since G is only determined up to a constant.

Case 2. General case. We look for a solution of the form $w = e^\lambda \sigma$. We find that the equation

$$(4) \quad w_{\bar{z}} = \mu w_z + \gamma w + \delta$$

is equivalent to

$$\sigma_{\bar{z}} + \lambda_{\bar{z}}\sigma = \mu\sigma_z + \mu\lambda_z\sigma + \gamma\sigma + e^{-\lambda}\delta,$$

or

$$\sigma(\lambda_{\bar{z}} - \mu\lambda_z - \gamma) + \sigma_{\bar{z}} = \mu\sigma_z + e^{-\lambda}\delta.$$

By Case 1, there are $C^{n+2+\alpha}$ functions λ, σ satisfying

$$\begin{aligned} \lambda_{\bar{z}} &= \mu\lambda_z + \gamma; & \lambda(0) &= 0, & \lambda_z(0) &= 0 \\ \sigma_{\bar{z}} &= \mu\sigma_z + e^{-\lambda}\delta; & \sigma(0) &= a, & \sigma_z(0) &= b. \end{aligned}$$

Then $w = e^\lambda \sigma$ is $C^{n+2+\alpha}$ and satisfies (4), and $w(0) = a$, $w_z(0) = b$. ❖

Notice that Proposition 29 does not give a C^∞ isothermal coordinate system in the C^∞ case; for although the equation $w_{\bar{z}} = \mu w_z$ will have $C^{n+1+\alpha}$ solutions for all n , these solutions might be defined on smaller and smaller neighborhoods of 0. But this is now easy to take care of. First let us note that if (u, v) is an isothermal coordinate system, and $f: \mathbb{C} \rightarrow \mathbb{C}$ is complex analytic, with f' never 0, then $f \circ (u, v)$ is also an isothermal coordinate system, since f is angle preserving. We can also prove this from our equation $w_{\bar{z}} = \mu w_z$, for since $f_{\bar{z}} = 0$, the chain rule gives

$$\begin{aligned}(f \circ w)_z &= (f_z \circ w) \cdot w_z \\ (f \circ w)_{\bar{z}} &= (f_z \circ w) \cdot w_{\bar{z}},\end{aligned}$$

and hence we have $(f \circ w)_{\bar{z}} = \mu \cdot (f \circ w)_z$. This argument can also be reversed, allowing us to prove

30. PROPOSITION. If μ is a $C^{n+\alpha}$ function with $|\mu| < 1$, and w is any solution of

$$(*) \quad w_{\bar{z}} = \mu w_z,$$

then w is $C^{n+1+\alpha}$. So if μ is C^∞ , any solution w is also C^∞ .

In particular, there is a C^∞ isothermal coordinate system around any point of a surface with a C^∞ metric.

PROOF. We know that around any point there is *some* $C^{n+1+\alpha}$ solution σ of $(*)$ which has an inverse around that point. So we can write

$$(1) \quad w = f \circ \sigma$$

for some f . Then the chain rule gives

$$\begin{aligned}w_z &= (f_z \circ \sigma) \cdot \sigma_z + (f_{\bar{z}} \circ \sigma) \bar{\sigma}_z \\ w_{\bar{z}} &= (f_z \circ \sigma) \cdot \sigma_{\bar{z}} + (f_{\bar{z}} \circ \sigma) \bar{\sigma}_{\bar{z}}.\end{aligned}$$

Since w is a solution of $(*)$, we have

$$(f_z \circ \sigma) \sigma_{\bar{z}} + (f_{\bar{z}} \circ \sigma) \bar{\sigma}_{\bar{z}} = \mu [(f_z \circ \sigma) \sigma_z + (f_{\bar{z}} \circ \sigma) \bar{\sigma}_z].$$

Since σ is a solution, this leads to

$$(2) \quad (f_{\bar{z}} \circ \sigma) [\bar{\sigma}_{\bar{z}} - \mu \bar{\sigma}_z] = 0.$$

Since $\sigma_{\bar{z}} = \mu\sigma_z$ implies that

$$\bar{\sigma}_z = \overline{(\sigma_{\bar{z}})} = \bar{\mu}\overline{(\sigma_z)} = \bar{\mu}\bar{\sigma}_{\bar{z}},$$

we see that

$$\begin{aligned}\bar{\sigma}_{\bar{z}} - \mu\bar{\sigma}_z &= \bar{\sigma}_{\bar{z}} - \mu\bar{\mu}\bar{\sigma}_{\bar{z}} \\ &= \bar{\sigma}_{\bar{z}}(1 - |\mu|^2).\end{aligned}$$

This is non-zero, since $|\mu| < 1$, and σ has non-zero Jacobian at the point in question. It follows from (2) that $f_{\bar{z}} = 0$, i.e., f is analytic. Then (1) shows that w must be $C^{n+1+\alpha}$ too. ♦

ADDENDUM 2

IMMERSED SPHERES WITH
CONSTANT MEAN CURVATURE

Let $f: U \rightarrow M$ be an immersion (for $U \subset \mathbb{R}^2$ open) which is conformal, so that the components E, F, G of I_f satisfy

$$E = G, \quad F = 0;$$

such immersions always exist by the results* of Addendum 1. From equation (B) on pg. III.136 we have

$$(1) \quad K = k_1 k_2 = \frac{ln - m^2}{E^2}$$

$$(2) \quad H = \frac{1}{2}(k_1 + k_2) = \frac{l + n}{2E}.$$

A little calculation shows that the Codazzi-Mainardi equations (pg. III.56) become

$$\begin{aligned} l_y - m_x &= \frac{E_y}{2E}(l + n) = E_y H \\ m_y - n_x &= -\frac{E_x}{2E}(l + n) = -E_x H. \end{aligned}$$

But

$$EH = \frac{l + n}{2} \implies \begin{cases} E_y H = -E H_y + \frac{l_y}{2} + \frac{n_y}{2} \\ E_x H = -E H_x + \frac{l_x}{2} + \frac{n_x}{2}, \end{cases}$$

so the Codazzi-Mainardi equations can be written

$$(3) \quad \begin{aligned} \left(\frac{l - n}{2}\right)_x + m_y &= E H_x \\ \left(\frac{l - n}{2}\right)_y - m_x &= -E H_y. \end{aligned}$$

* At present we need Proposition 29 or 30, but we could make do with the much simpler Theorem 18, since it follows from (hard) theorems on partial differential equations that a surface of constant mean curvature must be analytic (see pg. V.147).

If we define the function $\Phi: U \rightarrow \mathbb{C}$ by

$$(4) \quad \Phi = \frac{l-n}{2} - i \cdot m,$$

then

$$\begin{aligned} |\Phi|^2 &= \frac{(l-n)^2}{4} + m^2 = \frac{(l+n)^2}{4} + m^2 - ln \\ &= E^2(H^2 - K) \quad \text{by (1) and (2)} \\ &= E^2(k_1 - k_2)^2/4. \end{aligned}$$

Thus the umbilics on $f(U)$ are the image of the zeros of Φ . Notice that if H is constant, so that $H_x = H_y = 0$, then equations (3) are precisely the Cauchy-Riemann equations for Φ ; thus Φ is complex analytic. So we immediately have

31. LEMMA. If M is a connected surface immersed in \mathbb{R}^3 with constant mean curvature, then either all points of M are umbilics, or else the umbilics are isolated.

PROOF. Since the analytic function Φ is identically zero if its zeros have an accumulation point, we see that for every $p \in M$ one of two possibilities must hold:

- (1) p has a neighborhood with no umbilics, except perhaps p ,
- (2) p has a neighborhood all of whose points are umbilics.

But the set of points p satisfying (1) is open, and so is the set of points p satisfying (2). Since M is connected, either (1) holds everywhere, or (2) holds everywhere. ♦

Now consider the lines of curvature on M , or rather their images in U under the map f^{-1} . Formula (D) on pg. III.136 says that a vector $\mathbf{v} = (a_1, a_2)$ is tangent to one of these curves if and only if

$$\begin{aligned} 0 &= \det \begin{pmatrix} a_2^2 & -a_1 a_2 & a_1^2 \\ E & 0 & E \\ l & m & n \end{pmatrix} \\ &= -E[-ma_1^2 + (l-n)a_1 a_2 + ma_2^2] \\ &\quad \{l, m, n \text{ evaluated at the point where } \mathbf{v} \\ &\quad \text{is considered to be a tangent vector}\}. \end{aligned}$$

Thus \mathbf{v} is tangent to $[f^{-1} \text{ of}]$ a line of curvature if and only if

$$-m\{dx(\mathbf{v})\}^2 + (l - n) dx(\mathbf{v}) dy(\mathbf{v}) + m\{dy(\mathbf{v})\}^2 = 0.$$

We can write the left side of this equation as the imaginary part of a complex number, namely

$$\begin{aligned} \operatorname{Im} \left[\frac{l - n}{2} - i \cdot m \right] \cdot [\{dx(\mathbf{v})\}^2 - \{dy(\mathbf{v})\}^2 + 2i dx(\mathbf{v}) dy(\mathbf{v})] \\ = \operatorname{Im} \Phi \cdot [\{dx(\mathbf{v})\}^2 - \{dy(\mathbf{v})\}^2 + 2i dx(\mathbf{v}) dy(\mathbf{v})]. \end{aligned}$$

Introducing the complex-valued 1-form dz , as on page 319, we can thus write our equation as

$$\operatorname{Im} \Phi \cdot \{dz(\mathbf{v})\}^2 = 0.$$

For any complex number $w \neq 0$, we let $\arg w$ be some angle between the x -axis and the ray from 0 through w , so that $w = |w|e^{i \arg w}$. Then the above equation holds if and only if there is an integer m with

$$\begin{aligned} m\pi &= \arg \Phi \cdot \{dz(\mathbf{v})\}^2 \\ &= \arg \Phi + 2 \arg dz(\mathbf{v}), \end{aligned}$$

or

$$(*) \quad \arg dz(\mathbf{v}) = -\frac{1}{2} \arg \Phi + \frac{m\pi}{2} \quad \text{for some integer } m.$$

In a neighborhood of an isolated umbilic of our surface M with constant H we consider the 1-dimensional distribution Δ formed by the multiples of the principle vectors with the larger principal curvature, say. The index of this distribution was defined in Addendum 2 to Chapter 4. We can now compute it in terms of Φ .

32. PROPOSITION. Let $f: U \rightarrow M$ be a conformal immersion into a surface M of constant mean curvature H , with corresponding analytic function Φ . Suppose that $p = f(0)$ is an isolated umbilic, so that $\Phi(0) = 0$, and consequently

$$\Phi(z) = a_n z^n + \cdots \quad a_n \neq 0, \quad n \geq 1.$$

Then the index of Δ at p is $-n/2$.

PROOF. We consider the distribution on U which is f^{-1} of Δ . Let $c: [0, 1] \rightarrow U$ be a small circle around 0. To compute the index of the distribution at 0, we must choose a continuous function $\theta: [0, 1] \rightarrow \mathbb{R}$ such that $\theta(t)$ is an angle between the x -axis and the direction of the distribution at $c(t)$; then the index is $[\theta(1) - \theta(0)]/2\pi$. First choose a continuous function $\phi: [0, 1] \rightarrow \mathbb{R}$ such that $\phi(t)$ is an argument for $\Phi(c(t))$. Then equation (*) shows that we must have

$$\theta(t) = -\frac{1}{2}\phi(t) + \frac{m\pi}{2},$$

where the integer m must be constant, by continuity. So the index in question is

$$\frac{1}{2\pi}[\pi(1) - \theta(0)] = -\frac{1}{2} \cdot \frac{1}{2\pi}[\phi(1) - \phi(0)].$$

But standard complex analysis results say that $\phi(1) - \phi(0) = 2\pi n$. [Here is a direct proof. Clearly $[\phi(1) - \phi(0)]/2\pi$ is just the degree of the map α from S^1 to $\mathbb{C} - \{0\}$ defined by

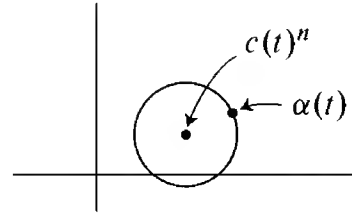
$$\begin{aligned} \alpha(t) &= \frac{1}{a_n} \Phi(c(t)) = c(t)^n [1 + \cdots] \\ &= c(t)^n [1 + d(t)], \end{aligned}$$

where we have

$$|d(t)| < 1 \quad \text{for a sufficiently small circle } c.$$

Now

$$|\alpha(t) - c(t)^n| = |c(t)^n d(t)| < |c(t)|^n.$$



So the line segment from $c(t)^n$ to $\alpha(t)$ does not contain 0. This means that α and $t \mapsto c(t)^n$ are homotopic as maps from S^1 to $\mathbb{C} - \{0\}$. So they have the same degree. But the degree of $t \mapsto c(t)^n$ is n .] ♦

All of this leads up to

33. THEOREM (H. HOPF). If M is an immersed sphere in \mathbb{R}^3 with constant mean curvature H , then M is a standard sphere.

PROOF. If all points of M were not umbilics, then by Lemma 31 there would be only finitely many umbilics. By Proposition 32, the index of Δ at each umbilic would be negative. This contradicts Theorem 4-20, since $\chi(M) = 2 > 0$. ♦

ADDENDUM 3

IMBEDDED SURFACES WITH
CONSTANT MEAN CURVATURE

In this Addendum we will prove that a compact *imbedded* surface $M \subset \mathbb{R}^3$ with constant mean curvature H_0 must be a standard sphere. Essentially the same proof works for imbedded hypersurfaces in \mathbb{R}^{n+1} , H^{n+1} , or an open hemisphere of S^{n+1} . The proof depends on a simple ingenious geometric construction, together with analytic results (Theorems 10-17 and 10-20) from Addendum 2 to Chapter 10; the proofs of these theorems can be read right now, for they do not depend on any material from Chapter 10 proper. These analytic results are applied to the present situation as follows.

Consider a surface given as the graph of a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$, and introduce the standard abbreviations

$$p = \frac{\partial h}{\partial x}, \quad q = \frac{\partial h}{\partial y}$$

$$r = \frac{\partial^2 h}{\partial x^2}, \quad s = \frac{\partial^2 h}{\partial x \partial y}, \quad t = \frac{\partial^2 h}{\partial y^2}.$$

For the condition that the surface has constant mean curvature H_0 we find, from (B') on pg. III.137, the equation

$$(*) \quad 0 = (1 + q^2)r - 2pq s + (1 + p^2)t - 2H_0(1 + p^2 + q^2)^{3/2}$$

$$= F(p, q, r, s, t).$$

Now let h_1 and h_2 be two solutions of (*), with corresponding partials p_1, \dots, t_1 and p_2, \dots, t_2 . If we denote the partial derivatives of F with respect to its 5 arguments as F_p, \dots, F_t , then at all points of \mathbb{R}^2 we have

$$0 = F(p_1, q_1, r_1, s_1, t_1) - F(p_2, q_2, r_2, s_2, t_2)$$

$$= \int_0^1 \frac{d}{d\tau} F(\tau p_1 + (1 - \tau)p_2, \dots, \tau t_1 + (1 - \tau)t_2) d\tau$$

$$= \int_0^1 (p_1 - p_2)F_p(\bullet) + \dots + (t_1 - t_2)F_t(\bullet) d\tau$$

where $\bullet = (\tau p_1 + (1 - \tau)p_2, \dots, \tau t_1 + (1 - \tau)t_2)$

$$= A \cdot (p_1 - p_2) + B \cdot (q_1 - q_2) + C \cdot (r_1 - r_2)$$

$$+ D \cdot (s_1 - s_2) + E \cdot (t_1 - t_2), \quad \text{say.}$$

Setting $u = h_1 - h_2$, and letting p, q, \dots, t now denote the partials of u , we see that u satisfies the equation

$$(**) \quad A \cdot p + B \cdot q + C \cdot r + D \cdot s + E \cdot t = 0.$$

34. LEMMA. Let h_1 and h_2 be two functions whose graphs are surfaces of the same constant mean curvature H_0 , both functions being defined either in a neighborhood of 0 in \mathbb{R}^2 , or in a neighborhood of 0 in the closed half-plane $\{(x, y) : y \geq 0\}$. Suppose that $h_1 \geq h_2$ in this domain, and that $h_1(0) = h_2(0)$. If h_1 and h_2 are defined only in the half-plane, assume also that $\partial h_1 / \partial x(0) = \partial h_2 / \partial x(0)$. Then $h_1 = h_2$ in a neighborhood of 0, or in a neighborhood of 0 in $\{(x, y) : y \geq 0\}$.

PROOF. Notice that for all $(\lambda, \mu) \neq (0, 0)$ we have

$$\begin{aligned} F_r \lambda^2 + F_s \lambda \mu + F_t \mu^2 &= (1 + q^2) \lambda^2 - 2pq \lambda \mu + (1 + p^2) \mu^2 \\ &= \lambda^2 + \mu^2 + (q\lambda - p\mu)^2 > 0, \end{aligned}$$

where F_r, F_s, F_t are evaluated at any point of \mathbb{R}^5 . So we also have

$$\begin{aligned} C \lambda^2 + D \lambda \mu + E \mu^2 &= \int_0^1 F_r(\bullet) \lambda^2 + F_s(\bullet) \lambda \mu + F_t(\bullet) \mu^2 d\tau \\ &> 0. \end{aligned}$$

Thus Theorems 10-17 and 10-20 apply to the solution $u = h_1 - h_2$ of equation (**). ♦

For the geometric part of the proof, we first note that the standard spheres are the only compact surfaces which have a plane of symmetry in every direction. In fact,

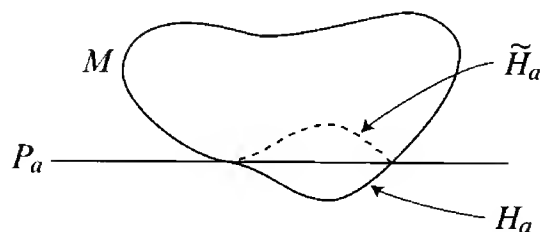
35. LEMMA. If $A \subset \mathbb{R}^3$ is bounded and has a plane of symmetry in every direction, then A is invariant under all rotations about some point $*$ (hence A is a union of concentric spheres around $*$).

PROOF. Choose 3 mutually orthogonal planes P_1, P_2, P_3 which are planes of symmetry for A , and let $*$ be the unique point in $P_1 \cap P_2 \cap P_3$. Let P be any other plane of symmetry. It is easy to see that if P does not go through $*$, then suitable compositions of the reflections through P_1, P_2, P_3 , and P will take any given point in A to points arbitrarily far from $*$. So if A is bounded, then we must have $*$ in P . Thus A is invariant under reflection through every plane through $*$. This implies that A is invariant under all rotations about $*$. ♦

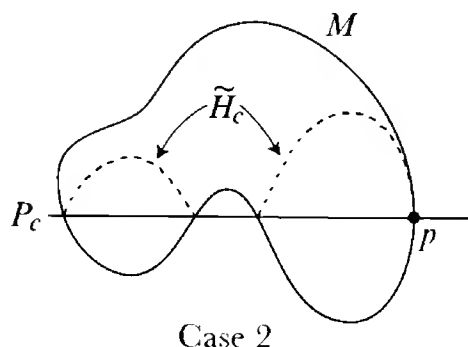
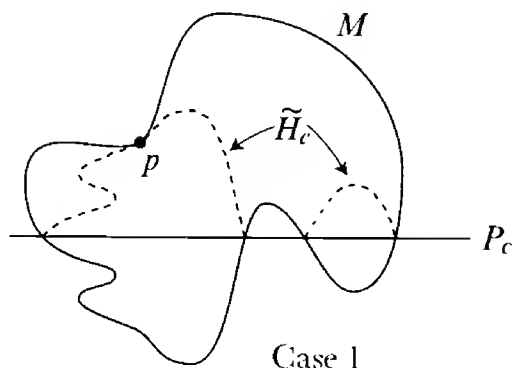
It is this symmetry property of spheres which we will establish for any surface of constant mean curvature.

36. THEOREM (ALEXANDROV). Let M be a compact surface imbedded in \mathbb{R}^3 with constant mean curvature H_0 . Then M has a plane of symmetry in every direction, so M is a standard sphere.

PROOF. We know (Theorem I.11-14) that M is the boundary of some closed domain D . We can assume that our direction is the z -axis, and that M is placed so that it lies in the region where $z \geq 0$, and touches the plane $z = 0$. For each $a > 0$, let P_a be the plane $z = a$. The set of points of M which lie below P_a is a “hump” H_a . Let \tilde{H}_a be the reflection of H_a in P_a . For sufficiently small



$a > 0$, the set \tilde{H}_a will lie inside D . Consider the set of all $b > 0$ such that \tilde{H}_a lies in D for $0 \leq a \leq b$. This set clearly has a largest element c . There are then two possible cases, as illustrated below.



In the first case, there is a point $p \in \tilde{H}_c \cap M$ which is not on P_c . From the definition of c , it is easy to see that near p the surfaces M and \tilde{H}_c are the graphs of two functions h_1, h_2 with $h_1 \geq h_2$. Then Lemma 35 shows that M and \tilde{H}_c coincide in a neighborhood of p .

If there is no point $p \in \tilde{H}_c \cap M - P_c$, then we must have the situation shown in the second figure: for some point $p \in P_c$, the surface M has a vertical tangent plane at p . The part of M which lies above or on P_c is a surface-with-boundary, and near p it can be represented as a function h_1 on a closed

half-plane perpendicular to the (x, y) -plane. Similarly, \tilde{H}_c can be represented as a function h_2 on the same closed half-plane. This time we have $h_2 \geq h_1$. Lemma 35 shows that \tilde{H}_c coincides near p with the part of M which lies above or on P_c .

Now let \tilde{K} be the component of \tilde{H}_c which contains the point p (in either Case 1 or Case 2). This component \tilde{K} is the reflection in P_c of a component $K \subset H_c$. The argument of the above two paragraphs, together with a simple connectedness argument, shows that $\tilde{K} \subset M$. But $\tilde{K} \cup K \subset M$ is already a compact manifold. So we must have $\tilde{K} \cup K = M$. Thus M is symmetric with respect to the plane P_c . ♦

One of the most interesting aspects of this proof is the fact that constancy of the mean curvature was used in such a weak way. There are numerous other conditions which can be treated similarly; Alexandrov has a whole series of papers on this subject. A somewhat later paper by Alexandrov [2] generalizes Theorem 36 so as to allow many types of self-intersections of M . For example, if $M \subset \mathbb{R}^3$ is a compact surface, bounding a domain D , and $f: M \rightarrow \mathbb{R}^3$ is an immersion which can be extended to an immersion of D into \mathbb{R}^3 , then $f(M)$ does not have constant mean curvature unless it is a standard sphere. Naturally, the counterexamples of Wente, Abresch, and Kapouleas (page 311) cannot be of this sort.

ADDENDUM 4

THE SECOND VARIATION OF VOLUME

In this Addendum we will derive the formula for the second variation of volume, and give some applications. The calculation itself is a real bitch, and even the final formula is quite involved, so some preliminaries will be required.

1. For a submanifold $M \subset N$ and a vector $\xi \in M_p^\perp$ we have the map $A_\xi: M_p \rightarrow M_p$ with

$$\langle A_\xi(X), Y \rangle = \langle s(X, Y), \xi \rangle.$$

Since A_ξ is symmetric, it has n real eigenvalues $\lambda_1, \dots, \lambda_n$. We will let

$$\Sigma_2(\xi) = \sum_{i=1}^n \lambda_i^2 = \text{trace } A_\xi^2.$$

If ξ denotes, instead, a section of the normal bundle $\text{Nor } M$, then $\Sigma_2(\xi)$ is a function on M .

2. Given $\xi \in M_p^\perp$, we define the “partial Ricci tensor”

$$\text{Ric}_M(\xi) = - \sum_{i=1}^n \langle R'(\xi, X_i) X_i, \xi \rangle,$$

where X_1, \dots, X_n is an orthonormal basis of M_p ; it is easily seen that this does not depend on the choice of X_1, \dots, X_n . Naturally, $\text{Ric}_M(\xi)$ denotes a function on M if ξ denotes a section of the normal bundle $\text{Nor } M$.

3. Recall from Addendum 1 of Chapter 7 that if W is a vector field tangent along the manifold M , then $\text{div } W$ is the function on M defined by

$$(\text{div } W)(p) = \text{trace}(X_p \mapsto \nabla_{X_p} W) = \sum_{i=1}^n \langle \nabla_{X_i} W, X_i \rangle,$$

where X_1, \dots, X_n is any orthonormal basis of M_p .

4. We also recall from this same Addendum that we have defined the Laplacian $\Delta\psi$ for a section ψ of a vector bundle over a Riemannian manifold M ; to define this we needed a connection on the bundle. For a submanifold $M \subset N$ of a Riemannian manifold N , we have the induced metric on M , and a connection D on the normal bundle defined by

$$D_X \psi = \perp \nabla'_X \psi, \quad \nabla' = \text{the covariant derivative in } N.$$

Thus if ψ is a section of the normal bundle, we have

$$\Delta\psi(p) = \sum_{j=1}^n \perp \nabla'_{X_j(p)} (\perp \nabla'_{X_j} \psi),$$

where X_1, \dots, X_n is an orthonormal moving frame with

$$\nabla_{X_i} X_j(p) = 0, \quad \nabla = \text{covariant derivative in } M.$$

5. We will require the following properties of contractions $X \lrcorner \omega$ and Lie derivatives $L_Z \omega$:

- | | | |
|-----|---|--|
| (a) | $Z \lrcorner (\phi \wedge \eta) = (Z \lrcorner \phi) \wedge \eta + (-1)^k \phi \wedge (Z \lrcorner \eta)$ | for ϕ a k -form
[Problem I.7-4] |
| (b) | $L_Z(\phi \wedge \eta) = L_Z \phi \wedge \eta + \phi \wedge L_Z \eta$ | } [Problem I.7-18] |
| (c) | $L_Z \omega = Z \lrcorner d\omega + d(Z \lrcorner \omega)$ | |
| (d) | $L_Z d\omega = dL_Z \omega$ | |
| (e) | $L_{Y+Z} \omega = L_Y \omega + L_Z \omega$ | [Problem I.5-14] |
| (f) | $L_Z(Y \lrcorner \omega) = [Z, Y] \lrcorner \omega + Y \lrcorner L_Z \omega$ | [an exercise, using
Problem I.7-18(c)]. |

6. Finally, there is one important way that the second variation formula for volume will differ from the second variation formula for energy. If $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$ is a variation of $\bar{\alpha}(0) = f: M \rightarrow N$, and $W(p) = \partial\alpha/\partial u(0, p)$ is the variation vector field, then our formula will involve not merely W , but also $\tilde{W} = \partial\alpha/\partial u$. We will define vector fields $\mathsf{T}\tilde{W}$ and $\perp\tilde{W}$ along α by writing $\tilde{W}(u, p) = \mathsf{T}\tilde{W}(u, p) + \perp\tilde{W}(u, p)$, where $\mathsf{T}\tilde{W}(u, p)$ is tangent to $\bar{\alpha}(u)(M)$ at $u(p)$ and $\perp\tilde{W}(u, p)$ is orthogonal to $\bar{\alpha}(u)(M)$ at $u(p)$.

37. THEOREM. Let $f: M \rightarrow N$ be a minimal immersion of an oriented n -dimensional manifold (-with-boundary) M into a Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$, and let $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$ be a variation of f through immersions. Let W be the variation vector field, and let $\tilde{W} = \partial\alpha/\partial u$. If $\Gamma(u)$ is the volume form on M determined by the metric $\bar{\alpha}(u)^*\langle \cdot, \cdot \rangle$ and the given orientation of M , then

$$\begin{aligned} \ddot{\Gamma}(0) = & [\text{Ric}_M(\perp W) - \Sigma_2(\perp W) - \langle \perp W, \Delta(\perp W) \rangle] \cdot \Gamma(0) \\ & + d(\text{div } \mathsf{T}W \cdot (\mathsf{T}W \lrcorner \Gamma(0)) + \mathsf{T}[\perp\tilde{W}, \mathsf{T}\tilde{W}] \lrcorner \Gamma(0)). \end{aligned}$$

PROOF. We will regard this proof as a continuation of the proof of Theorem 11; we will refer to equations (1)–(10) in that proof, and therefore commence our numbering of equations with (11). Once again we first consider a point $p_0 \in M$ where $W(p_0)$ is not tangent to $f(M)$. We choose V as before, and assume that f is just the inclusion $i: V \rightarrow N$. We will use all the notation introduced in the proof of Theorem 11, and we will also introduce the abbreviations

$$(11) \quad \theta^j = i^* \phi^j \quad 1 \leq j \leq n.$$

Since our immersion is minimal ($\eta = 0$), equation (9) shows that

$$(12) \quad i^*\{Z \lrcorner d\Phi\} = 0 \quad \text{for all vector fields } Z \text{ along } V.$$

Using (6), we can write $d\Phi$ as

$$(13) \quad d\Phi = \sum_{r=n+1}^m \phi^r \wedge \mu_r, \quad \text{for } \mu_r = \sum_{j=1}^n \phi^1 \wedge \cdots \wedge \psi_r^j \wedge \cdots \wedge \phi^n.$$

Then (12) becomes

$$\sum_{r=n+1}^m \theta^r(Z) i^* \mu_r = 0, \quad \text{using (a) on page 356 and } i^* \phi^r = 0.$$

Since this is true for arbitrary Z , we have

$$(14) \quad i^* \mu_r = 0, \quad \text{and hence} \quad \sum_{j=1}^n \psi_r^j(X_j) = 0 \quad \text{along } V.$$

Now let us apply equation (5) to all u , not just $u = 0$. We obtain

$$\dot{\Gamma}(u) = \bar{\alpha}(u)^*(\tilde{W} \lrcorner d\Phi) + \bar{\alpha}(u)^*d(\tilde{W} \lrcorner \Phi).$$

As before, this implies that

$$(15) \quad \begin{aligned} \ddot{\Gamma}(0) &= i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} + i^*\{L_{\tilde{W}}d(\tilde{W} \lrcorner \Phi)\} \\ &= i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} + d(i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner \Phi)\}) \quad \text{by (d).} \end{aligned}$$

Once again we will show that the two terms on the right are precisely the terms appearing in the statement of the theorem.

The first term is the one that will give us all the difficulties, and we will use some preliminary tricks to make the calculations manageable. First of all, we

want to be more particular in our choice of the moving frame $X_1, \dots, X_n, X_{n+1}, \dots, X_m$. We will assume that $X_1(p_0), \dots, X_n(p_0)$ is a basis of eigenvectors for $A_{\perp W(p_0)}$, with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$. This means that

$$\begin{aligned}
 \lambda_j \delta_{jk} &= \langle A_{\perp W(p_0)} X_j(p_0), X_k(p_0) \rangle \\
 &= \langle s(X_k(p_0), X_j(p_0)), \perp W(p_0) \rangle \\
 &= \langle \perp \nabla'_{X_k} X_j, \perp W \rangle && \text{at } p_0 \\
 &= \left\langle \sum_{r=n+1}^m \psi_j^r(X_k) X_r, \perp W \right\rangle && \text{at } p_0 \\
 &= - \sum_{r=n+1}^m \phi^r(W) \cdot \psi_r^j(X_k) && \text{at } p_0,
 \end{aligned}$$

and consequently,

$$(16) \quad \sum_{r=n+1}^m \phi^r(W) i^* \psi_r^j = -\lambda_j \theta^j \quad \text{at } p_0.$$

We still have considerable leeway in the choice of our moving frame X_1, \dots, X_m . We can replace it with a new moving frame $\bar{X}_1, \dots, \bar{X}_m$ defined by

$$\bar{X}_\alpha = \sum_{\beta=1}^m M_\alpha^\beta X_\beta,$$

where (M_α^β) is a matrix of functions such that

- (i) (M_α^β) is always orthogonal, $(M_\alpha^\beta)^{-1} = (M_\beta^\alpha)$,
- (ii) $M_r^j = M_j^r = 0 \quad 1 \leq j \leq n, \quad n+1 \leq r \leq m$,
- (iii) $(M_\alpha^\beta(p_0)) = I$.

Condition (i) means that the new moving frame is orthonormal, and condition (ii) implies that (1) and (2) still hold, so that the \bar{X}_j are tangent to the $\bar{\alpha}(u)(V)$, while the \bar{X}_r are orthogonal. Condition (iii) means that the frame is not changed at p_0 , so that equation (16) still holds. The dual 1-forms $\bar{\phi}^\alpha$ are related to the ϕ^β by

$$\bar{\phi}^\alpha = \sum_{\beta=1}^m M_\alpha^\beta \phi^\beta. \quad \bar{\phi}^\alpha(p_0) = \phi^\alpha(p_0).$$

so the corresponding connection forms $\bar{\psi}_\beta^\alpha$ satisfy

$$\begin{aligned}
 -\sum_{\beta=1}^m \bar{\psi}_\beta^\alpha \wedge \bar{\phi}^\beta &= d\bar{\phi}^\alpha = \sum_{\beta=1}^m dM_\alpha^\beta \wedge \phi^\beta + \sum_{\beta=1}^m M_\alpha^\beta \wedge d\phi^\beta \\
 &= \sum_{\beta=1}^m dM_\alpha^\beta \wedge \phi^\beta - \sum_{\beta,\gamma=1}^m M_\alpha^\beta \psi_\gamma^\beta \wedge \phi^\gamma \\
 &= \sum_{\beta=1}^m \sum_{\delta=1}^m M_\beta^\delta dM_\alpha^\beta \wedge \bar{\phi}^\delta - \sum_{\beta,\gamma,\delta=1}^m M_\alpha^\beta M_\delta^\gamma \psi_\gamma^\beta \wedge \bar{\phi}^\delta \\
 &= -\sum_{\beta=1}^m \left[\sum_{\gamma,\delta=1}^m M_\alpha^\delta M_\beta^\gamma \psi_\gamma^\delta - \sum_{\delta=1}^m M_\delta^\beta dM_\alpha^\delta \right] \wedge \bar{\phi}^\beta.
 \end{aligned}$$

Now $\sum_{\gamma,\delta} M_\alpha^\delta M_\beta^\gamma \psi_\gamma^\delta$ is easily seen to be skew-symmetric with respect to α and β , since $\psi_\gamma^\delta = -\psi_\delta^\gamma$. Since (M_α^β) is orthogonal and $M(p_0) = I$, we also have skew-symmetry for $\sum_{\delta} M_\delta^\beta dM_\alpha^\delta$ at p_0 . So Proposition II.7-4 (which is really a result about forms on a single vector space) shows that at p_0 we have

$$(iv) \quad \bar{\psi}_\beta^\alpha(p_0) = \psi_\beta^\alpha(p_0) - dM_\alpha^\beta(p_0).$$

Now we claim that it is possible to choose M_α^β so that

$$(v) \quad \begin{cases} dM_k^j(p_0) = \psi_k^j(p_0) & 1 \leq j, k \leq n \\ dM_s^r(p_0) = \psi_s^r(p_0) & n+1 \leq r, s \leq m. \end{cases}$$

In fact, for every unit tangent vector X at p_0 we can define

$$\begin{aligned}
 M_k^j(\exp tX) &= \exp(t\psi_k^j(X)) & 1 \leq j, k \leq n \\
 M_s^r(\exp tX) &= \exp(t\psi_s^r(X)) & n+1 \leq r, s \leq m \\
 M_r^j &= M_j^r = 0,
 \end{aligned}$$

where the exp on the right is the ordinary exponential of matrices. Then the matrices (M_β^α) satisfy (v); they also satisfy (i)–(iii), the matrices (M_k^j) and (M_s^r) being orthogonal since they are exp of skew-symmetric matrices. In connection with (iv), we thus see that our moving frame can be picked so that it satisfies not only (16), but also

$$(17) \quad \begin{cases} \psi_k^j(p_0) = 0 & 1 \leq j, k \leq n \\ \psi_s^r(p_0) = 0 & n+1 \leq r, s \leq m. \end{cases}$$

In addition to this special choice of the moving frame, we require another preliminary move. We are trying to show that $i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\}$ is the first term in the formula of the theorem. We notice that this term involves only the perpendicular component $\perp W$ of W . We can ease the strain of the calculations by first proving that the same is true of the expression that we have to work with. In the following lemma, $\top \tilde{W}$ and $\perp \tilde{W}$ actually denote extensions of these vector fields to a neighborhood of image α .

38. LEMMA. For a minimal immersion we have

$$i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} = i^*\{L_{\perp \tilde{W}}(\perp \tilde{W} \lrcorner d\Phi)\}.$$

PROOF. By property (c), which we stated before the theorem, we have

$$i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} = i^*\{\tilde{W} \lrcorner d(\tilde{W} \lrcorner d\Phi)\}.$$

For vector fields Y and Z in N , define

$$\mathcal{J}(Y, Z) = i^*\{Y \lrcorner d(Z \lrcorner d\Phi)\},$$

so that

$$i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} = \mathcal{J}(\tilde{W}, \tilde{W}).$$

It is clear that \mathcal{J} is bilinear over \mathbb{R} . We will show that $\mathcal{J}(Y, Z) = 0$ if either Y or Z is tangent along M . The lemma then follows by writing $\tilde{W} = \top \tilde{W} + \perp \tilde{W}$.

Suppose first that Y is tangent along M . Then

$$\begin{aligned} \mathcal{J}(Y, Z) &= i^*\{Y \lrcorner d(Z \lrcorner d\Phi)\} \\ &= i^*\{L_Y(Z \lrcorner d\Phi) - d(Y \lrcorner Z \lrcorner d\Phi)\} \quad \text{by (c).} \end{aligned}$$

Now equation (12) tells us that $Z \lrcorner d\Phi$ gives 0 when applied to tangent vectors of M . The same is clearly true for $Y \lrcorner Z \lrcorner d\Phi$, since Y is itself tangent along M . So

$$i^*\{Y \lrcorner Z \lrcorner d\Phi\} = 0, \quad \text{and hence} \quad i^*\{d(Y \lrcorner Z \lrcorner d\Phi)\} = 0.$$

On the other hand, since Y is tangent to M we clearly also have

$$i^*\{L_Y(Z \lrcorner d\Phi)\} = 0.$$

Thus $\mathcal{J}(Y, Z) = 0$.

Now suppose that Z is tangent along M . We have

$$\begin{aligned}
 \mathcal{L}(Y, Z) &= i^*\{Y \lrcorner d(Z \lrcorner d\Phi)\} \\
 &= i^*\{Y \lrcorner L_Z d\Phi\} && \text{by (c)} \\
 &= i^*\{[Y, Z] \lrcorner d\Phi\} + i^*\{L_Z(Y \lrcorner d\Phi)\} && \text{by (f)} \\
 &= i^*\{[Y, Z] \lrcorner d\Phi\} \\
 &\quad + i^*\{Z \lrcorner d(Y \lrcorner d\Phi)\} + i^*\{d(Z \lrcorner Y \lrcorner d\Phi)\} && \text{by (c)} \\
 &= i^*\{[Y, Z] \lrcorner d\Phi\} + \mathcal{L}(Z, Y) + i^*\{d(Z \lrcorner Y \lrcorner d\Phi)\}.
 \end{aligned}$$

The first term is 0 by (12). The second term is 0 by what we have already proved. The third term is 0 for the same reason that $i^*\{d(Y \lrcorner Z \lrcorner d\Phi)\}$ was 0 before. **Q.E.D.**

We are now finally ready to carry out the computation.

Step 1. We claim that

$$(18) \quad i^*\{L_{\widetilde{W}}\phi^r\} = 0 = i^*\{L_{\perp\widetilde{W}}\phi^r\}, \quad n+1 \leq r \leq m.$$

To see this, choose Y to be a vector field tangent to V and let $i_*Y = X$. Then

$$\begin{aligned}
 i^*\{L_{\widetilde{W}}\phi^r\}(Y) &= L_{\widetilde{W}}\phi^r(X) \\
 &= d(W \lrcorner \phi^r)(X) + (W \lrcorner d\phi^r)(X) && \text{by (c)} \\
 &= X(\phi^r(W)) + d\phi^r(W, X) \\
 &= X(\phi^r(W)) \\
 &\quad + [W(\phi^r(X)) - X(\phi^r(W)) - \phi^r([W, X])] && \text{by pg. I.215} \\
 &= -\phi^r([W, X]).
 \end{aligned}$$

But if t^1, \dots, t^n is a coordinate system around p_0 in V , then X is a linear combination of $\partial\alpha/\partial t^1, \dots, \partial\alpha/\partial t^n$,

$$X = \sum_{j=1}^n a_j \frac{\partial\alpha}{\partial t^j}.$$

We have

$$\left[W, \frac{\partial\alpha}{\partial t^j}\right] = \left[\frac{\partial\alpha}{\partial u}, \frac{\partial\alpha}{\partial t^j}\right] = \alpha_* \left(\left[\frac{\partial}{\partial u}, \frac{\partial}{\partial t^j}\right]\right) = 0,$$

so

$$[W, X] = \left[W, \sum_{j=1}^n a_j \frac{\partial\alpha}{\partial t^j}\right] = - \sum_{j=1}^n W(a_j) \frac{\partial\alpha}{\partial t^j} \quad \text{by pg. I.156.}$$

Thus $[W, X]$ is also tangent to V , so $\phi^r([W, X]) = 0$. This shows that $i^*\{L_{\tilde{W}}\phi^r\} = 0$.

We also have

$$i^*\{L_{\top\tilde{W}}\phi^r\} = -\phi^r([\top W, X]) = 0,$$

since $[\top W, X]$ is tangent to V . Hence $i^*\{L_{\perp\tilde{W}}\phi^r\} = 0$ also.

Step 2. For $1 \leq j \leq n$, we have

$$\begin{aligned} i^*\{L_{\tilde{W}}\phi^j\} &= i^*\{d(\tilde{W} \lrcorner \phi^j)\} + i^*\{\tilde{W} \lrcorner d\phi^j\} && \text{by (c)} \\ &= d(\phi^j(W)) - i^*\left\{\tilde{W} \lrcorner \sum_{\alpha=1}^m \psi_{\alpha}^j \wedge \phi^{\alpha}\right\} \\ &= d(\phi^j(W)) - \sum_{k=1}^m \psi_k^j(W)\theta^k + \sum_{\alpha=1}^m \phi^{\alpha}(W)i^*\psi_{\alpha}^j && \text{by (a)}. \end{aligned}$$

Using (16) and (17) we see that

$$(19) \quad i^*\{L_{\tilde{W}}\phi^j\} = d(\phi^j(W)) - \lambda_j\theta^j \quad \text{at } p_0.$$

Similarly, we find that

$$(20) \quad i^*\{L_{\perp\tilde{W}}\phi^j\} = -\lambda_j\theta^j \quad \text{at } p_0.$$

Step 3. Using the second structural equation to express $d\psi_r^j$ in terms of the curvature forms Ψ_r^j , we have

$$\begin{aligned} i^*\{L_{\perp\tilde{W}}\psi_r^j\} &= i^*\{\perp\tilde{W} \lrcorner d\psi_r^j\} + i^*\{d(\perp\tilde{W} \lrcorner \psi_r^j)\} && \text{by (c)} \\ &= -i^*\left\{\perp\tilde{W} \lrcorner \sum_{\gamma=1}^m \psi_{\gamma}^j \wedge \psi_r^{\gamma}\right\} + i^*\{\perp\tilde{W} \lrcorner \Psi_r^j\} + d(\psi_r^j(\perp W)). \end{aligned}$$

Because of equation (17), each term $\psi_{\gamma}^j \wedge \psi_r^{\gamma}$ always has one factor equal to 0 at p_0 , so we obtain

$$(21) \quad i^*\{L_{\perp\tilde{W}}\psi_r^j\} = i^*\{\perp\tilde{W} \lrcorner \Psi_r^j\} + d(\psi_r^j(\perp W)) \quad \text{at } p_0.$$

Step 4. Referring to (13) for the definition of μ_r , we now compute

$$\begin{aligned} i^*\{L_{\perp\tilde{W}}\mu_r\} &= i^*\left\{L_{\perp\tilde{W}} \sum_{j=1}^n \phi^1 \wedge \cdots \wedge \psi_r^j \wedge \cdots \wedge \phi^n\right\} \\ &= \sum_{j=1}^n \theta^1 \wedge \cdots \wedge i^*\{L_{\perp\tilde{W}}\psi_r^j\} \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{j=1}^n \left[\sum_{k \neq j} \theta^1 \wedge \cdots \wedge i^*\{L_{\perp\tilde{W}}\phi^k\} \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n \right]. \end{aligned}$$

Substituting from (20) and (21), and rearranging slightly, we have

$$\begin{aligned} i^*\{L_{\perp\tilde{W}}\mu_r\} &= \sum_{j=1}^n \theta^1 \wedge \cdots \wedge i^*\{\perp\tilde{W} \lrcorner \Psi_r^j\} \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{j=1}^n \left[\sum_{k \neq j} \theta^1 \wedge \cdots \wedge -\lambda_k \theta^k \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n \right] \\ &\quad + \sum_{j=1}^n \theta^1 \wedge \cdots \wedge d(\psi_r^j(\perp W)) \wedge \cdots \wedge \theta^n \quad \text{at } p_0. \end{aligned}$$

Notice that

$$\begin{aligned} \sum_{j=1}^n \left[\sum_{k \neq j} \theta^1 \wedge \cdots \wedge -\lambda_k \theta^k \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n \right] \\ = \sum_{j=1}^n \left(\sum_{k \neq j} -\lambda_k \right) \theta^1 \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n. \end{aligned}$$

But $\sum_{k=1}^n \lambda_k = 0$, since our immersion is minimal; so $\sum_{k \neq j} -\lambda_k = \lambda_j$. Thus

$$\begin{aligned} (22) \quad i^*\{L_{\perp\tilde{W}}\mu_r\} &= \sum_{j=1}^n \theta^1 \wedge \cdots \wedge i^*\{\perp\tilde{W} \lrcorner \Psi_r^j\} \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{j=1}^n \lambda_j \theta^1 \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{j=1}^n \theta^1 \wedge \cdots \wedge d(\psi_r^j(\perp W)) \wedge \cdots \wedge \theta^n \quad \text{at } p_0. \end{aligned}$$

Step 5. We have

$$\begin{aligned} i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} &= i^*\{L_{\perp\tilde{W}}(\perp\tilde{W} \lrcorner d\Phi)\} && \text{by Lemma 38} \\ &= i^*\{\perp\tilde{W} \lrcorner d(\perp\tilde{W} \lrcorner d\Phi)\} && \text{by (c)} \\ &= i^*\{\perp\tilde{W} \lrcorner L_{\perp\tilde{W}}d\Phi\} && \text{by (c) again} \\ &= \sum_{r=n+1}^m i^*\{\perp\tilde{W} \lrcorner L_{\perp\tilde{W}}(\phi^r \wedge \mu_r)\} && \text{by (13)} \\ &= \sum_{r=n+1}^m i^*\{\perp\tilde{W} \lrcorner (L_{\perp\tilde{W}}\phi^r \wedge \mu_r)\} \\ &\quad + \sum_{r=n+1}^m i^*\{\perp\tilde{W} \lrcorner (\phi^r \wedge L_{\perp\tilde{W}}\mu_r)\} && \text{by (b).} \end{aligned}$$

When we expand the first of these sums by (a), we obtain two terms, one involving $i^*\mu_r$ and one involving $i^*\{L_{\perp\tilde{W}}\phi^r\}$. These will both be 0, by (14) and (18), so we obtain

$$\begin{aligned} i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} &= \sum_{r=n+1}^m i^*\{\perp\tilde{W} \lrcorner (\phi^r \wedge L_{\perp\tilde{W}}\mu_r)\} \\ &= \sum_{r=n+1}^m \phi^r(W) i^*\{L_{\perp\tilde{W}}\mu_r\} \quad \text{by (a).} \end{aligned}$$

Substituting in from (22), we obtain

$$\begin{aligned} (23) \quad i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner d\Phi)\} &= \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \theta^1 \wedge \cdots \wedge i^*\{\perp\tilde{W} \lrcorner \Psi_r^j\} \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \lambda_j \theta^1 \wedge \cdots \wedge i^*\psi_r^j \wedge \cdots \wedge \theta^n \\ &\quad + \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \theta^1 \wedge \cdots \wedge d(\psi_r^j(\perp W)) \wedge \cdots \wedge \theta^n \quad \text{at } p_0 \\ &= S_1 + S_2 + S_3, \quad \text{say.} \end{aligned}$$

Step 6. We will see what each of these sums gives when applied to the n -tuple of vectors $X_1(p_0), \dots, X_n(p_0)$.

Recall that

$$\Psi_r^j(X_\alpha, X_\beta) = \langle R'(X_\alpha, X_\beta)X_r, X_j \rangle.$$

Thus we have

$$\begin{aligned} S_1(X_1, \dots, X_n) &= \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \Psi_r^j(\perp W, X_j) && \text{at } p_0 \\ &= \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \langle R'(\perp W, X_j)X_r, X_j \rangle && \text{at } p_0 \\ &= - \sum_{r=n+1}^m \phi^r(W) \langle R'(\perp W, X_j)X_j, X_r \rangle && \text{at } p_0 \\ &= -\langle R'(\perp W, X_j)X_j, \perp W \rangle = \text{Ric}_M(\perp W) && \text{at } p_0. \end{aligned}$$

Hence

$$(24) \quad S_1 = \text{Ric}_M(\perp W) \cdot \Gamma(0) \quad \text{at } p_0.$$

Next we have

$$\begin{aligned} S_2(X_1, \dots, X_n) &= \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n \lambda_j \psi_r^j(X_j) \quad \text{at } p_0 \\ &= \sum_{j=1}^n \lambda_j \sum_{r=n+1}^m \phi^r(W) \psi_r^j(X_j) \quad \text{at } p_0 \\ &= - \sum_{j=1}^n \lambda_j^2 \quad \text{by (16).} \end{aligned}$$

Hence

$$(25) \quad S_2 = -\Sigma_2(\perp W) \cdot \Gamma(0) \quad \text{at } p_0.$$

To evaluate S_3 , we note that $d(\psi_r^j(\perp W)) = \sum_i d(\psi_r^j(\perp W))(X_i) \cdot \theta^i$. So we obtain

$$\begin{aligned} (26) \quad S_3(X_1, \dots, X_n) &= \sum_{r=n+1}^m \phi^r(W) \sum_{j=1}^n d(\psi_r^j(\perp W))(X_j) \quad \text{at } p_0 \\ &= \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) X_j(\psi_r^j(\perp W)) \quad \text{at } p_0. \end{aligned}$$

Step 7. The coefficient of $\Gamma(0)$ in the statement of the theorem will clearly be completely accounted for as soon as we show that

$$(27) \quad S_3(X_1, \dots, X_n) = -\langle \perp W, \Delta(\perp W) \rangle \quad \text{at } p_0.$$

Equation (17) implies that $\langle \nabla'_{X_i} X_k, X_j \rangle = 0$ at p_0 , and hence that $\nabla_{X_i} X_k = 0$ at p_0 , where ∇ is the covariant derivative in V . So

$$\Delta(\perp W) = \sum_{j=1}^n \perp \nabla'_{X_j} (\perp \nabla'_{X_j} \perp W).$$

Now

$$\begin{aligned}
 \nabla'_{X_j} \perp W &= \nabla'_{X_j} \left(\sum_{r=n+1}^m \phi^r(W) X_r \right) \\
 &= \sum_{r=n+1}^m X_j(\phi^r(W)) X_r + \sum_{r=n+1}^m \phi^r(W) \nabla'_{X_j} X_r \\
 &= \sum_{r=n+1}^m X_j(\phi^r(W)) X_r + \sum_{r=n+1}^m \phi^r(W) \sum_{\alpha=1}^m \psi_r^\alpha(X_j) X_\alpha,
 \end{aligned}$$

so

$$\perp(\nabla'_{X_j} \perp W) = \sum_{r=n+1}^m [X_j(\phi^r(W)) + \sum_{s=n+1}^m \phi^s(W) \psi_s^r(X_j)] X_r.$$

Hence

$$\begin{aligned}
 \nabla'_{X_j} (\perp \nabla'_{X_j} \perp W) &= \sum_{r=n+1}^m X_j \left(X_j(\phi^r(W)) + \sum_{s=n+1}^m \phi^s(W) \psi_s^r(X_j) \right) \cdot X_r \\
 &\quad + \sum_{r=n+1}^m [X_j(\phi^r(W)) + \sum_{s=n+1}^m \phi^s(W) \psi_s^r(X_j)] \cdot \sum_{\alpha=1}^m \psi_r^\alpha(X_j) X_\alpha.
 \end{aligned}$$

Using (17), we obtain*

$$\perp \nabla'_{X_j} (\perp \nabla'_{X_j} \perp W) = \sum_{r=n+1}^m [X_j(X_j(\phi^r(W))) + \sum_{s=n+1}^m \phi^s(W) X_j(\psi_s^r(X_j))] X_r \text{ at } p_0,$$

and therefore

$$\begin{aligned}
 (28) \quad \langle \perp W, \Delta(\perp W) \rangle &= \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) \cdot [X_j(X_j(\phi^r(W))) \\
 &\quad + \sum_{s=n+1}^m \phi^s(W) X_j(\psi_s^r(X_j))] \quad \text{at } p_0 \\
 &= \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) X_j(X_j(\phi^r(W))) \quad \text{at } p_0.
 \end{aligned}$$

since $\psi_s^r = -\psi_r^s$.

*Note that $X_j(\psi_s^r(X_j))$ need not be zero at p_0 , even though $\psi_s^r(X_j) = 0$ at p_0 .

We can find out something about the $X_j(\phi^r(W))$ by writing (18) in the form

$$\begin{aligned}
 0 &= i^*\{L_{\perp \tilde{W}} \phi^r\} = i^*\{d(\perp \tilde{W} \lrcorner \phi^r)\} + i^*\{\perp \tilde{W} \lrcorner d\phi^r\} && \text{by (c)} \\
 &= d(\phi^r(W)) - i^*\left\{\perp \tilde{W} \lrcorner \sum_{\alpha=1}^m \psi_\alpha^r \wedge \phi^\alpha\right\} \\
 &= d(\phi^r(W)) - \sum_{k=1}^n \psi_k^r(\perp W) \theta^k + \sum_{s=n+1}^m \phi^s(W) i^* \psi_s^r && \text{by (a).}
 \end{aligned}$$

This gives us

$$X_j(\phi^r(W)) = -\psi_r^j(\perp W) - \sum_{s=n+1}^m \phi^s(W) \psi_s^r(X_j);$$

using (17) we obtain

$$X_j(X_j(\phi^r(W))) = -X_j(\psi_r^j(\perp W)) - \sum_{s=n+1}^m \phi^s(W) X_j(\psi_s^r(X_j)) \quad \text{at } p_0.$$

Substituting into (28), we get

$$\begin{aligned}
 \langle \perp W, \Delta(\perp W) \rangle &= - \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) X_j(\psi_r^j(\perp W)) \\
 &\quad - \sum_{j=1}^n \sum_{r,s=n+1}^m \phi^r(W) \phi^s(W) X_j(\psi_s^r(X_j)) \\
 &= - \sum_{j=1}^n \sum_{r=n+1}^m \phi^r(W) X_j(\psi_r^j(\perp W)) \quad \text{at } p_0, \text{ since } \psi_s^r = -\psi_r^s.
 \end{aligned}$$

This proves (27), and completes our calculation of the first term in (15).

The second term in (15) will not be nearly so bad. We have

$$\begin{aligned}
 (29) \quad i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner \Phi)\} &= i^*\{\tilde{W} \lrcorner d(\tilde{W} \lrcorner \Phi)\} && \text{by (c)} \\
 &= i^*\{\tilde{W} \lrcorner L_{\tilde{W}} \Phi\} && \text{by (c)} \\
 &= i^*\left\{\tilde{W} \lrcorner \sum_{j=1}^n \phi^1 \wedge \cdots \wedge L_{\tilde{W}} \phi^j \wedge \cdots \wedge \phi^n\right\} && \text{by (b).}
 \end{aligned}$$

To show that

$$(30) \quad i^*\{L_{\tilde{W}}(\tilde{W} \lrcorner \Phi)\} = \operatorname{div} \mathbf{T}W \cdot (\mathbf{T}W \lrcorner \Gamma(0)) + \mathbf{T}[\perp \tilde{W}, \mathbf{T}\tilde{W}] \lrcorner \Gamma(0),$$

it obviously suffices to show that both sides give the same result when applied to any $(n-1)$ -tuple $(X_1, \dots, \widehat{X_l}, \dots, X_n)$ at p_0 . Since the X_i 's enter symmetrically, we can assume, by renumbering, that $l = 1$. Now (29) gives

$$\begin{aligned}
 i^*\{L_{\widetilde{W}}(\widetilde{W} \sqcup \Phi)\}(X_2, \dots, X_n) &= \left(\sum_{j=1}^n \phi^1 \wedge \dots \wedge L_{\widetilde{W}} \phi^j \wedge \dots \wedge \phi^n \right)(W, X_2, \dots, X_n) \quad \text{at } p_0 \\
 &= (L_{\widetilde{W}} \phi^1 \wedge \dots \wedge \phi^n)(W, X_2, \dots, X_n) \\
 &\quad + \sum_{j=2}^n (\phi^1 \wedge \dots \wedge L_{\widetilde{W}} \phi^j \wedge \dots \wedge \phi^n)(W, X_2, \dots, X_n) \quad \text{at } p_0.
 \end{aligned}$$

In computing the first term, the only permutations of (W, X_2, \dots, X_n) that do not give zero are interchanges of W with one X_j ; in the second sum only the given order (W, X_2, \dots, X_n) produces a non-zero result. So

$$\begin{aligned}
 i^*\{L_{\widetilde{W}}(\widetilde{W} \sqcup \Phi)\}(X_2, \dots, X_n) &= \left[(L_{\widetilde{W}} \phi^1)(W) - \sum_{j=2}^n \phi^j(W)(L_{\widetilde{W}} \phi^1)(X_j) \right] \\
 &\quad + \sum_{j=2}^n \phi^1(W)(L_{\widetilde{W}} \phi^j)(X_j) \quad \text{at } p_0 \\
 &= (L_{\widetilde{W}} \phi^1)(W) - \sum_{j=1}^n \phi^j(W)(L_{\widetilde{W}} \phi^1)(X_j) \\
 &\quad + \phi^1(W) \sum_{j=1}^n (L_{\widetilde{W}} \phi^j)(X_j) \quad \text{at } p_0 \\
 &\quad \text{[since } j = 1 \text{ gives the same new term in each sum]} \\
 &= (L_{\widetilde{W}} \phi^1)(W) - (L_{\widetilde{W}} \phi^1) \left(\sum_{j=1}^n \phi^j(W) X_j \right) \\
 &\quad + \phi^1(W) \sum_{j=1}^n (L_{\widetilde{W}} \phi^j)(X_j) \quad \text{at } p_0 \\
 &= (L_{\widetilde{W}} \phi^1)(\perp W) + \phi^1(W) \sum_{j=1}^n [X_j(\phi^j(W)) - \lambda_j] \quad \text{at } p_0 \\
 &\quad \text{[by (19)]}
 \end{aligned}$$

$$\begin{aligned}
&= (L_{\tilde{W}}\phi^1)(\perp W) + \phi^1(W) \sum_{j=1}^n X_j(\phi^j(\top W)) && \text{at } p_0 \\
&\quad [\text{since } \sum_j \lambda_j = 0] \\
&= (W \lrcorner d\phi^1)(\perp W) + d(\tilde{W} \lrcorner \phi^1)(\perp W) \\
&\quad + \phi^1(W) \sum_{j=1}^n X_j(\langle \top W, X_j \rangle) && \text{at } p_0 \\
&= d\phi^1(W, \perp W) + \perp W(\phi^1(\tilde{W})) \\
&\quad + \phi^1(W) \sum_{j=1}^n [\langle \nabla_{X_j} \top W, X_j \rangle + \langle \top W, \nabla_{X_j} X_j \rangle] && \text{at } p_0 \\
&= [W(\phi^1(\perp \tilde{W})) - \perp W(\phi^1(\tilde{W})) - \phi^1([\tilde{W}, \perp \tilde{W}])] + \perp W(\phi^1(\tilde{W})) \\
&\quad + \phi^1(W) \sum_{j=1}^n \langle \nabla_{X_j} \top W, X_j \rangle && \text{at } p_0 \\
&\quad [\text{since } \sum_j \nabla_{X_j} X_j = \eta = 0] \\
&= -\phi^1([\tilde{W}, \perp \tilde{W}]) + \phi^1(W) \sum_{j=1}^n \langle \nabla_{X_j} \top W, X_j \rangle \\
&= \langle \top[\perp \tilde{W}, \top \tilde{W}], X_1 \rangle + \langle \top W, X_1 \rangle \sum_{j=1}^n \langle \nabla_{X_j} \top W, X_j \rangle.
\end{aligned}$$

This is exactly the value of the right side of (30) on X_2, \dots, X_n ; we have thus completed the calculation of the second term in (15).

Finally, we again dispose of the general case, where $W(p_0)$ may be tangent to V , by considering $\mathbf{N} = N \times \mathbb{R}$, with the product metric, and the map $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow \mathbf{N}$ defined by

$$\alpha(u, p) = (\alpha(u, p), u).$$

We recall that

$$\mathbf{W}(p) = (W(p), 1) \quad \text{and} \quad \boldsymbol{\eta}(p) = (\eta, 0).$$

So $\tilde{\alpha}(0)$ is minimal if $\bar{\alpha}(0)$ is. If \mathbf{R}' is the curvature tensor in \mathbf{N} , then we have

$$\begin{aligned}
\text{Ric}_M(\perp W) &= \text{Ric}_M((\perp W, 1)) \\
&= - \sum_{i=1}^n \langle \mathbf{R}'(\perp W, X_i) X_i, \perp W \rangle - \sum_{i=1}^n \langle \mathbf{R}'(1, X_i) X_i, 1 \rangle,
\end{aligned}$$

for X_1, \dots, X_n an orthonormal basis of M . Using the fact that we have a product metric, we easily find that

$$\begin{aligned}\operatorname{Ric}_M(\perp \mathbf{W}) &= - \sum_{i=1}^n \langle R'(\perp W, X_i) X_i, \perp W \rangle \\ &= \operatorname{Ric}_M(\perp W).\end{aligned}$$

The map $\mathbf{s}(p): M_p \times M_p \rightarrow M_p^\perp$ is obviously given by

$$\mathbf{s}(p)(X, Y) = (s(X, Y), 0), \quad X, Y \in M_p,$$

so the map $\mathbf{A}_{\perp \mathbf{W}}$ is given by

$$\langle \mathbf{A}_{\perp \mathbf{W}}(X), Y \rangle = \langle (s(X, Y), 0), (\perp W, 1) \rangle = \langle s(X, Y), \perp W \rangle = \langle A_{\perp W}(X), Y \rangle.$$

Consequently,

$$\Sigma_2(\perp \mathbf{W}) = \Sigma_2(\perp W).$$

We also have

$$\Delta(\perp \mathbf{W}) = \Delta((\perp W, 1)) = (\Delta(\perp W), 0),$$

and hence

$$\langle \perp \mathbf{W}, \Delta(\perp \mathbf{W}) \rangle = \langle (\perp W, 1), (\Delta(\perp W), 0) \rangle = \langle \perp W, \Delta(\perp W) \rangle.$$

Since we obviously have $\mathbf{T}\mathbf{W} = \mathbf{T}W$, we have

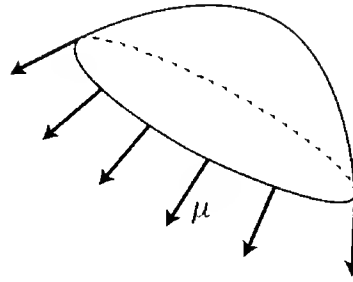
$$\operatorname{div} \mathbf{T}\mathbf{W} \cdot (\mathbf{T}\mathbf{W} \lrcorner \Gamma(0)) = \operatorname{div} \mathbf{T}W \cdot (\mathbf{T}W \lrcorner \Gamma(0)).$$

Finally,

$$\begin{aligned}[\perp \tilde{\mathbf{W}}, \mathbf{T}\tilde{\mathbf{W}}] &= [(\perp \tilde{W}, 1), (\mathbf{T}\tilde{W}, 0)] \\ &= [\perp \tilde{W}, \mathbf{T}\tilde{W}] + [1, \mathbf{T}\tilde{W}] \\ &= [\perp \tilde{W}, \mathbf{T}\tilde{W}],\end{aligned}$$

the second bracket vanishing since there is clearly a coordinate system x^1, \dots, x^m, τ on \mathbf{N} with $\partial/\partial x^1 = \mathbf{T}\tilde{W}$ and $\partial/\partial \tau = 1$. Thus the result for α implies the result for α . ♦

To integrate the result of Theorem 37 succinctly, we introduce the **outward pointing unit normal of ∂M along f** (see the picture on the next page): for each $q \in \partial M$, we define $\mu(q) \in N_{f(q)}$ to be the unit vector tangent to $f(M)$, perpendicular to $f(\partial N)$, and outward pointing. Recall (pg. I.260) that the orientation for ∂M is chosen so that v_1, \dots, v_{n-1} is positively oriented at q if and only if $\mu(q), v_1, \dots, v_{n-1}$ is positively oriented on M .



39. COROLLARY. Let $f: M \rightarrow N$ be a minimal immersion of a compact oriented n -dimensional manifold-with-boundary M into a Riemannian manifold $(N^m, \langle \cdot, \cdot \rangle)$, and let $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow N$ be a variation of f through immersions. Let W be the variation vector field, let $\tilde{W} = \partial\alpha/\partial u$, and let μ be the outward pointing unit normal of ∂M along f . If $V(\bar{\alpha}(u))$ is the n -dimensional volume of M determined by the metric $\bar{\alpha}(u)^*\langle \cdot, \cdot \rangle$ and the given orientation of M , then

$$\begin{aligned} \left. \frac{d^2 V(\bar{\alpha}(u))}{du^2} \right|_{u=0} &= \int_M [\text{Ric}_M(\perp W) - \Sigma_2(\perp W) - \langle \perp W, \Delta(\perp W) \rangle] dV \\ &\quad + (-1)^{n+1} \int_{\partial M} [\text{div } \top W \cdot \langle \top W, \mu \rangle + \langle \top[\perp \tilde{W}, \top \tilde{W}], \mu \rangle] dV^{n-1}, \end{aligned}$$

where dV is the volume element determined by $f^*\langle \cdot, \cdot \rangle$, and dV^{n-1} is the induced volume element on ∂M . In particular, if α is a variation keeping ∂M fixed, then

$$\left. \frac{d^2 V(\bar{\alpha}(u))}{du^2} \right|_{u=0} = \int_M [\text{Ric}_M(\perp W) - \Sigma_2(\perp W) - \langle \perp W, \Delta(\perp W) \rangle] dV.$$

PROOF. Left to the reader. ♦

Problem 3 shows what our formula reduces to in the case of a geodesic $\gamma: [a, b] \rightarrow N$. Here we will consider the case of a hypersurface $M \subset N$, with $i: M \rightarrow N$ the inclusion map. Then $\perp W = hv$ for some function h , where v is a unit normal vector field. Since

$$\begin{aligned} \perp \nabla'_{X_j(p)}(\perp \nabla'_{X_j} hv) &= \perp \nabla'_{X_j(p)}(X_j(h)v) \\ &= X_j(p)(X_j(h)) \cdot v, \end{aligned}$$

we see that

$$\langle \perp W, \Delta(\perp W) \rangle = h \Delta h.$$

where Δ now denotes the Laplacian on functions, computed by means of the induced metric $i^*\langle \cdot, \cdot \rangle$ on M . So if Σ_2 denotes the sum of the squares of the

eigenvalues of the symmetric map $\Pi: M_p \times M_p \rightarrow \mathbb{R}$, then our integral becomes

$$\int_M [h^2 \operatorname{Ric}_M(v) - h^2 \Sigma_2 - h \Delta h] dV.$$

Suppose in particular, that we consider the variation by parallel surfaces, $\alpha(u, p) = \exp_p u \cdot v(p)$. Then $h = 1$ and (Problem 3-12) $\tilde{W}(u, p)$ is always perpendicular to $\tilde{\alpha}(u)(M)$; so the integral over ∂M drops out, and we obtain

$$\left. \frac{d^2 V(\tilde{\alpha}(u))}{du^2} \right|_{u=0} = \int_M [\operatorname{Ric}_M(v) - \Sigma_2] dV.$$

If N has sectional curvatures ≥ 0 , then $\operatorname{Ric}_M(v) \leq 0$, so we obtain

$$\left. \frac{d^2 V(\tilde{\alpha}(u))}{du^2} \right|_{u=0} \leq 0.$$

Moreover, we have strict inequality unless $\Sigma_2 = 0$, which happens only when $s = 0$, so that our hypersurface is totally geodesic. Thus a non-totally geodesic minimal hypersurface in a space of non-negative sectional curvature always has *greater* volume than nearby parallel surfaces.

Now let us consider a minimal immersion $f: M \rightarrow \mathbb{R}^3$, where M is a compact surface-with-boundary. Let $\alpha: (-\varepsilon, \varepsilon) \times M \rightarrow \mathbb{R}^3$ be a variation of f keeping ∂M fixed, such that $W = hN$ for some function h vanishing on ∂M , where N is a unit normal field. Then our formula becomes

$$\begin{aligned} (1) \quad \left. \frac{d^2 A(\tilde{\alpha}(u))}{du^2} \right|_{u=0} &= \int_M [-h^2(k_1^2 + k_2^2) - h \Delta h] dA \\ &\quad \text{where } k_1 \text{ and } k_2 = -k_1 \text{ are} \\ &\quad \text{the principal curvatures} \\ &= \int_M [2h^2 K - h \Delta h] dA \\ &= \int_M [2h^2 K + \mathbf{I}_f(\operatorname{grad} h, \operatorname{grad} h)] dA, \\ &\quad \text{by Proposition 7-59.} \end{aligned}$$

In particular, consider a compact 2-dimensional manifold-with-boundary $D \subset \mathbb{R}^2$ and a minimal immersion $\Phi: D \rightarrow \mathbb{R}^3$ given by

$$(*) \quad \begin{cases} \Phi^1 = \operatorname{Re} \int \frac{1}{2} F(w)(1 - w^2) dw \\ \Phi^2 = \operatorname{Re} \int \frac{i}{2} F(w)(1 + w^2) dw \\ \Phi^3 = \operatorname{Re} \int F(w)w dw \end{cases}$$

for a nowhere 0 complex analytic function $F: D \rightarrow \mathbb{C}$. For this immersion we have (page 274)

$$(2) \quad I_f = \Phi^*\langle \cdot, \cdot \rangle = \mu(dx \otimes dx + dy \otimes dy),$$

$$\text{where } \mu(z) = \frac{|F(z)|^2(1 + |z|^2)^2}{4}, \quad z = x + iy.$$

We can compute (Problem 5) that the curvature K for the metric $\Phi^*\langle \cdot, \cdot \rangle$ on D is given by

$$(3) \quad K(z) = \frac{-16}{|F(z)|^2(1 + |z|^2)^4}, \quad z = x + iy.$$

Suppose now that we have a variation α of Φ which keeps ∂D fixed, and such that $W(\Phi(z)) = h(z) \cdot N(z)$ for some function $h: D \rightarrow \mathbb{R}$ with $h = 0$ on ∂D . Using (2), we compute, from the last equation in the proof of Proposition 5, that

$$(4) \quad I_f(\text{grad } h, \text{grad } h)(z) = \frac{(h_1^2 + h_2^2)(z)}{\left(\frac{|F(z)|^2(1 + |z|^2)^2}{4}\right)}.$$

Substituting (4) and (3) into (1), and remembering that the volume element dA of I on D is

$$\sqrt{\det(g_{ij})} \, dx \wedge dy = \mu \, dx \wedge dy,$$

we obtain

$$\left. \frac{d^2 A(\bar{\alpha}(u))}{du^2} \right|_{u=0} = \int_D \left[[h_1(x, y)]^2 + [h_2(x, y)]^2 - \frac{8[h(x, y)]^2}{(1 + x^2 + y^2)^2} \right] dx \, dy.$$

Notice that this expression does not involve the original map $(*)$ at all; it involves only the region D , and the function h . If we recall (page 274) that $N = \sigma^{-1}$, we see that D contains the unit disc $B = \{(x, y) : x^2 + y^2 \leq 1\}$ if and only if the normal map N of Φ covers the whole southern hemisphere of the unit sphere.

40. THEOREM (SCHWARZ-RADO). If the interior of D contains the unit disc $B = \{(x, y) : x^2 + y^2 \leq 1\}$, then there is a function $h: D \rightarrow \mathbb{R}$ with $h = 0$ on ∂D such that

$$(1) \quad \int_D \left[h_1^2 + h_2^2 - \frac{8h^2}{(1 + x^2 + y^2)^2} \right] dx \, dy < 0.$$

Consequently, for every nowhere 0 complex analytic function $F: D \rightarrow \mathbb{C}$, the minimal surface $\Phi(D)$ given by $(*)$ does *not* have minimum area among all nearby surfaces with the same boundary.

(Since the solution to the Plateau problem tells us that there is *some* minimal disc with the same boundary as $\Phi(S^1)$, this proves that $\Phi(S^1)$ is the boundary of at least 2 different minimal surfaces.)

PROOF. Let $B(r) = \{(x, y) : x^2 + y^2 \leq r^2\}$, and define $h^r : B(r) \rightarrow \mathbb{R}$ by

$$(2) \quad h^r(x, y) = \frac{x^2 + y^2 - r^2}{x^2 + y^2 + r^2}.$$

Set

$$(3) \quad I(r) = \int_{B(r)} \left[(h^r_1)^2 + (h^r_2)^2 - \frac{8(h^r)^2}{(1 + x^2 + y^2)^2} \right] dx dy.$$

Substituting (2) into (3), we obtain the explicit formula

$$I(r) = \int_{B(r)} \frac{16(x^2 + y^2)r^4}{(x^2 + y^2 + r^2)^4} dx dy - \int_{B(r)} \frac{8(x^2 + y^2 - r^2)^2}{(x^2 + y^2 + r^2)^2(x^2 + y^2 + 1)^2} dx dy.$$

Making the substitution $x = u \cdot r$, $y = v \cdot r$, we get

$$I(r) = \int_B \frac{16(u^2 + v^2)}{(u^2 + v^2 + 1)^4} du dv - \int_B \frac{8(u^2 + v^2 - 1)^2 r^2}{(u^2 + v^2 + 1)^2(u^2 r^2 + v^2 r^2 + 1)^2} du dv.$$

Finally, computing $I'(1)$ by Leibniz's Rule, we obtain

$$I'(1) = 16 \int_B \frac{(u^2 + v^2 - 1)^3}{(u^2 + v^2 + 1)^5} du dv < 0.$$

On the other hand, we claim that $I(1) = 0$. To prove this, we use Proposition 7-59 and the fact that $h^r = 0$ on $\partial B(r)$ to write (3) as

$$I(r) = - \int_{B(r)} h^r \left[h^r_{11} + h^r_{22} + \frac{8h^r}{(1 + x^2 + y^2)^2} \right] dx dy;$$

then we just compute that the term in brackets is 0 for h^1 .

Since $I(1) = 0$ and $I'(1) < 0$, there is a number $r_0 > 1$ such that $I(r) < 0$ for $1 < r < r_0$. Now there is some r with $1 < r < r_0$ such that $D \supset B(r)$. Define h on D by

$$h(x, y) = \begin{cases} h^r(x, y) & (x, y) \in B(r) \\ 0 & \text{otherwise.} \end{cases}$$

This h has all the desired properties, except that the first partial derivatives of h are discontinuous on $B(r)$. However, it is easy to see that we can round off h to a C^∞ function without changing the sign of the integral in (1). ♦

PROBLEMS

1. Show that formula (*) on page 274 gives

$$\text{a catenoid for } F(w) = \frac{1}{w^2}$$

$$\text{a helicoid for } F(w) = \frac{i}{w^2}$$

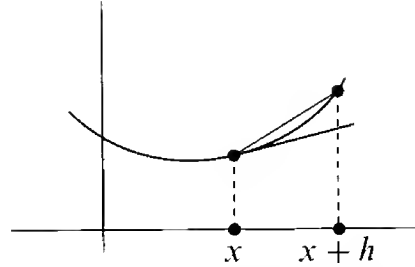
$$\text{Scherk's minimal surface for } F(w) = \frac{4}{1-w^4}.$$

2. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be convex.

(a) We have

$$f'(x^+) = \inf_{h>0} \frac{f(x+h) - f(x)}{h}$$

$$f'(x^-) = \sup_{h>0} \frac{f(x+h) - f(x)}{h}.$$



(b) If $f'(x)$ exists for all x , then f' is continuous. *Hint:* Consider $h > 0$, say, with $[f(x+h) - f(x)]/h < f'(x) + \varepsilon$.

3. Let $\gamma: [a, b] \rightarrow N$ be an arclength parameterized geodesic, with unit tangent vector $V = d\gamma/dt$, and let $\alpha: (-\varepsilon, \varepsilon) \times [a, b] \rightarrow V$ be a variation, with variation vector field \tilde{W} .

(a) If Z is a vector field along γ with $\langle V, Z \rangle = 0$, then $\perp \nabla'_V Z = \nabla'_V Z$.

(b) $\Delta(\perp W) = D^2 \perp W / dt^2$.

(c) We have

$$\left. \frac{d^2 L(\tilde{\alpha}(u))}{du^2} \right|_{u=0} = \int_a^b - \left\langle \frac{D^2 \perp W}{dt^2}, \perp W(t) \right\rangle - \langle R'(W, V)V, W \rangle(t) dt$$

$$+ \langle \nabla_V \top W, V \rangle \cdot \langle \top W, V \rangle + \langle [\perp W, \top W], V \rangle \Big|_a^b.$$

(d) Let

$$B = \langle \nabla_V \top W, V \rangle \cdot \langle \top W, V \rangle + \langle [\perp W, \top W], V \rangle$$

$$= \langle \nabla_V \top W, \top W \rangle + \langle \nabla_{\perp W} \top W, V \rangle - \langle \nabla_{\top W} \perp W, V \rangle.$$

Noting that $\top W$ is a multiple of V , say $\top W = hV$, show that

$$\langle \nabla_{\top W} \perp W, V \rangle = 0 \quad \text{and} \quad \langle \nabla_V \top W, \top W \rangle = \langle \nabla_{\top W} \top W, V \rangle.$$

Thus

$$\begin{aligned} B &= \langle \nabla_{\mathbf{T}W} \mathbf{T}W, V \rangle + \langle \nabla_{\perp W} \mathbf{T}W, V \rangle = \langle \nabla_W \mathbf{T}W, V \rangle \\ &= \langle \nabla_W W, V \rangle - \langle \nabla_W \perp W, V \rangle \\ &= \langle \nabla_W W, V \rangle + \langle \nabla_V \perp W, \perp W \rangle. \end{aligned}$$

(e) Conclude that

$$\begin{aligned} \frac{d^2 L(\tilde{\alpha}(u))}{du^2} \Big|_{u=0} &= \int_a^b \left\langle \frac{D\perp W}{dt}, \frac{D\perp W}{dt} \right\rangle - \langle R'(W, V)V, W \rangle(t) dt \\ &\quad + \langle \nabla_W W, V \rangle \Big|_a^b. \end{aligned}$$

4. If $M \subset \mathbb{R}^3$ is a minimal surface, then at any point $p \in M$ the Gaussian curvature $K(p)$ is given by

$$K(p) = -\frac{\langle \nu_* X, \nu_* X \rangle}{\langle X, X \rangle} \quad \text{for any } X \in M_p.$$

Hint: The numerator is $\text{III}(X, X)$.

5. Consider a minimal immersion $\Phi: V \rightarrow \mathbb{R}^3$ given by (*) on page 274, so that $N = \sigma^{-1}$ and $g_{ij} = \mu \delta_{ij}$, where

$$\mu(z) = \frac{|F(z)|^2(1 + |z|^2)^2}{4}.$$

Use Problems 4 and 7-20 to show that

$$\begin{aligned} K(z) &= -\left(\frac{2}{1 + |z|^2}\right)^2 \Big/ \mu(z) \\ &= \frac{-16}{|F(z)|^2(1 + |z|^2)^4}. \end{aligned}$$

6. (a) Let $M \subset \mathbb{R}^3$ be a minimal surface with $K < 0$ everywhere, and consider an imbedding $f: U \rightarrow M$ whose parameter lines are lines of curvature. Using the formulas on pg. III.217, show that if $k_1 > 0$ is the positive principal curvature, then

$$E(s, t) = S(s)/k_1(s, t), \quad G(s, t) = T(t)/k_1(s, t)$$

for certain functions $S, T > 0$. Then show that there is a new imbedding with

$$E = G = \frac{1}{k_1}.$$

Conclude that if $\langle \cdot, \cdot \rangle$ is the metric on M , then

$$\sqrt{-K} \langle \cdot, \cdot \rangle$$

is a flat metric (Ricci).

(b) Let $\langle \cdot, \cdot \rangle$ be a metric on a 2-dimensional manifold M such that $K < 0$ and $\sqrt{-K} \langle \cdot, \cdot \rangle$ is flat. Thus there is a coordinate system (u, v) such that

$$\begin{aligned} \sqrt{-K} \langle \cdot, \cdot \rangle = du \otimes du + dv \otimes dv &\implies \langle \cdot, \cdot \rangle = \frac{1}{\sqrt{-K}} (du \otimes du + dv \otimes dv) \\ &= g(du \otimes du + dv \otimes dv), \text{ say.} \end{aligned}$$

Using the formula on pg. III.217, show that

$$K = -\frac{1}{2g} \left[\left(\frac{g_v}{g} \right)_v + \left(\frac{g_u}{g} \right)_u \right].$$

Then show that there is an imbedding $f: U \rightarrow \mathbb{R}^3$ with

$$\begin{aligned} E = G &= \frac{1}{\sqrt{-K}}, & F &= 0 \\ l &= 1, & n &= -1, & m &= 0. \end{aligned}$$

Thus $f(U)$ is a minimal surface isometric to M .

MINI-BIBLIOGRAPHY FOR VOLUME IV

Brackets [] indicate journal articles, braces { } indicate books.

Abresch, U.

- [1] *Constant mean curvature tori in terms of elliptic functions*, J. Reine Angew. Math. **374** (1987), 169–192 (MR 88e:53006).

Alexandrov, A. D.

- [1] *Uniqueness theorems for surfaces in the large. I*, Vestnik Leningrad. Univ. **11** (1956), no. 19, 5–17 (Russian); Amer. Math. Soc. Transl. (2) **21** (1962), 341–354.
- [2] *A characteristic property of spheres*, Ann. Mat. Pura Appl. **58** (1962), 303–315.

Bianchi, L.

- [1] *Sulle superficie a curvatura nulla in geometria ellittica*, Ann. Mat. Pura Appl. **24** (1896), 93–129.

Blaschke, W.

- {1} *Vorlesungen über Differential Geometrie*, Vols. I and II, 3rd. ed., Chelsea, New York, 1967.
- {2} *Kreis und Kugel*, De Gruyter & Co., Berlin, 1966.

Boys, C. V.

- {1} *Soap Bubbles, Their Colors and the Forces Which Mold Them*, 3rd ed., Dover, New York, 1959.

Carmo, M. do and Lima, E.

- [1] *Isometric immersions with semi-definite second quadratic forms*, Arch. Math. (Basel) **20** (1969), 173–175.
- [2] *Immersions of manifolds with non-negative sectional curvatures*, Bol. Soc. Brasil. Mat. **2** (1972), 9–22.

Carmo, M. do and Warner, F. W.

- [1] *Rigidity and convexity of hypersurfaces in spheres*, J. Differential Geometry **4** (1970), 133–144.

Chern, S. S. and Lashof, R. K.

- [1] *On the total curvature of immersed manifolds*, Amer. J. Math. **79** (1957), 306–318; *II*, Michigan Math. J. **5** (1958), 5–12.

Courant, R.

- [1] *Soap film experiments with minimal surfaces*, Amer. Math. Monthly **47** (1940), 167–174.

Eisenhart, L. P.

- {1} *Riemannian Geometry*, Princeton University Press, Princeton, N.J., 1925.

Fialkow, A.

- [1] *Hypersurfaces of a space of constant curvature*, Ann. of Math. **39** (1938), 762–785.

Gromoll, D., Klingenberg, W., and Meyer, W.

- {1} *Riemannsche Geometrie im Grossen*, Lecture Notes in Math. No. 55, Springer-Verlag, Berlin-Heidelberg, 1968.

Hartshorne, R.

- {1} *Foundations of Projective Geometry*, W. A. Benjamin, New York, 1967.

Hopf, H.

- [1] *Über Flächen mit einer Relation zwischen den Hauptkrümmungen*, Math. Nachr. **4** (1951), 232–249.

Kapouleas, N.

- [1] *Compact constant mean curvature surfaces in Euclidean three-space*, J. Differential Geom. **33** (1991), no. 3., 683–715 (MR 93a:53007b).
- [2] *Constant mean curvature surfaces constructed by fusing Wente tori*, Proc. Nat. Acad. Sci. U.S.A. **89** (1992), no. 12, 5695–5698 (MR 93h:53011).

Klingenberg, W.

- {1} *Riemannian Geometry*, 2nd ed., Walter de Gruyter & Co, Berlin, 1995 (MR 95m:53003).

Lawson, H. B.

- [1] *Complete minimal surfaces in S^3* , Ann. of Math. (2) **92** (1970), 335–374.

Milnor, J.

- {2} *Morse Theory*, Ann. Math. Studies No. 51, Princeton University Press, Princeton, N. J., 1963.

O’Neil, B.

- [1] *Isometric immersions which preserve curvature operators*, Proc. Amer. Math. Soc. **13** (1962), 759–763.

Ryan, P. J.

- [1] *Homogeneity and some curvature conditions for hypersurfaces*, Tôhoku Math. J. (2) **21** (1969), 363–388.

Sacksteder, R.

- [1] *On hypersurfaces with no negative sectional curvatures*, Amer. J. Math. **82** (1960), 609–630.

Scheffers, G. W.

- {1} *Anwendung der Differential- und Integral-Rechnung auf Geometrie*, 2 vols., Veit & Co., Leipzig, 1901–1902.

Stoker, J. J.

- [1] *Über die Gestalt der positiv gekrümmten offenen Flächen im dreidimensionalen Raume*, Composito Math. **3** (1936), 55–89.

Warner, F. W.

- {1} *Foundations of Differentiable Manifolds and Lie Groups*, Scotts, Foresman and Co., Glenview, Illinois, 1971.

Wente, H. C.

- [1] *Counterexample to a conjecture of H. Hopf*, Pacific J. Math. **121** (1986), no. 1, 193–243 (MR 87d:53013).

NOTATION INDEX

CHAPTER 7A

B^n	7
H^n	2
\mathcal{H}^n	13
$H^n(K_0)$	2
$O^1(n+1)$	2
S^n	1
$S^n(K_0)$	1
$\phi: S^{n+} \rightarrow \mathbb{R}^n$	17
$\langle \cdot, \cdot \rangle$	2
$\langle \cdot, \cdot \rangle_p$	2

CHAPTER 7B

\mathbf{v}_i	21–24
κ_i	21–24

CHAPTER 7C

$A_{\xi p}$	34
$D_{X_p} \xi$	34
$F(E)$	52
$\text{Hom}(TM \times TM,$ Nor $M)$	37
Nor M	33
$O(E)$	52
$O(TM, E)$	54
R_D	41
s_{ij}^r	55
β_r^s	35
θ	52
$\varpi: \text{Nor } M \rightarrow M$	33
$\sigma(X)$	52
ϕ^α	54
Ψ_β^α	54
$\bar{\Psi}_s^r$	55
ψ_β^α	54
$\bar{\psi}_s^r$	55
Ω	54
ω	52

Π^r	35
$\Pi^r * \Pi^s$	38
$\tilde{\nabla}$	37

CHAPTER 7D

H	65
K	65
K_i	65
$K_{i;\xi}$	71
\mathcal{R}	69
\tilde{R}	66
ε	69
$\varepsilon^{j_1 \dots j_n}$	68
$\eta(p)$	71
σ_i	65
Λ	69

CHAPTER 7E

K_{ext}	86
K_{int}	86
$\mathbf{t}, \mathbf{u}, \mathbf{v}$	88
κ_g	89
κ_n	89
τ_g	89

CHAPTER 7F

$h: S^3 \rightarrow S^2$	107
\times	98

CHAPTER 7G

Ric	120
-----	-----

CHAPTER 7, ADDENDUM 1

$\text{div } X$	128, 129
$\text{grad } f$	128, 129
Δf	128, 132
$\Delta \xi$	138
$\Delta \langle \cdot, \cdot \rangle$	137
∇	128
∇^2	128

CHAPTER 7, ADDENDUM 2

\tilde{A}	141
G	145
H^k	145
h^k	145
Δ	143
δ	142
(ω, η)	143
$\langle \omega, \eta \rangle$	140–142
$*$	139

CHAPTER 7, ADDENDUM 3

$\Delta_1(f)$	152
$\Delta_1(f, g)$	152
$\Theta^2(u, v)$	153

CHAPTER 7, ADDENDUM 4

$A_{\xi_p}^k(X_p)$	171
$A^k(\xi_p; X_p)$	172
$D_{X_p}^k \xi$	172
$D^{[k]}_X \xi$	180, 181
\mathcal{F}_k	176
ℓ	166
$\text{Nor}^k M$	168
$\text{Nor}^k M_p$	167
$\text{Osc}^k M$	166
$\text{Osc}^k M_p$	166
$R^{[k]}$	180
s^k	168
\mathbf{s}^k	170
\mathbf{T}^k	167
$\mathbf{T}^{[k]}$	180
\perp^k	167
$\nabla'(X_1, \dots, X_k)$	166
$\#(M)$	166

CHAPTER 7, PROBLEMS

$v_1 \times \cdots \times v_{m-1}$	194
(z, z_1, z_2, z_3)	190

CHAPTER 8

$C(p)$	248
$\tilde{C}(p)$	248
$E(p)$	253
$E_*(W)$	202
$E_{**}(W_1, W_2)$	207
\tilde{M}	258
\mathbb{R}^*	251
$S(M)$	251
W_1, W_2	205
$\mu: S(M) \rightarrow \mathbb{R}^*$	251
Ω_γ	203

CHAPTER 9

$A: \mathcal{C}(X) \rightarrow \mathbb{R}$	304
$\text{Con}(X)$	304
$\mathcal{C}(X)$	303
$d(x, C)$	303
$J_{f*}(W)$	295, 298
$J_*(W)$	296, 298
$L: \text{Con}(X) \rightarrow \mathbb{R}$	304
$V_\varepsilon(C)$	303
$\dot{\Gamma}(u)$	286
$\rho(C_1, C_2)$	303
$\Phi_{(f,g)}$	270

CHAPTER 9, ADDENDUM 1

C^α	318
$C^{n+\alpha}$	318
$d\bar{z}$	319
$d\bar{z}$	319
$H(R, \alpha)$	336
$H(R, \alpha + 1)$	337

w_z	319	CHAPTER 9, ADDENDUM 4	
$w_{\bar{z}}$	319		
$\ w\ _R$	336	$\mathrm{Ric}_M(\xi)$	355
$\ \ w\ \ _R$	337	$\Sigma_2(\xi)$	355
$\partial w / \partial z$	319	$\top \tilde{W}$	356
$\partial w / \partial \bar{z}$	319	$\perp \tilde{W}$	356

INDEX

- Abresch, U., 311
- Alexandrov, A. D., 310, 354
- Alexandrov's Theorem, 353
- Allendoerfer, C. B., 163
- Analytic flat surfaces in \mathbb{R}^3 , 118
- Area of an immersed surface, 259
- Associated minimal surfaces, 275
- Asymptotic directions, 89

- Beltrami, E., 133; *see also* Laplace-Beltrami operator
- Beltrami equations, 319
- Beltrami's Theorem, 19
- Beltrami-Enneper Theorem, 89
- Berger, M., 239
- Bernstein's Theorem, 267
- Bianchi, L., 94
- Bianchi identity, second, 182
- Bilinear function, index of, 3
- Blaschke, W., 303, 309, 311, 313
- Bochner, S., 150
- Bochner's Lemma, 134
- Bonnet, O., 229, 307
- Boys, C. V., 281
- Bubbles, soap, 310
- Bundle
 - isomorphism, 48
 - normal, 33
 - sphere, 251
- Burstin, C., 163

- Calculus of variations
 - direct methods in, 241
 - in several variables, 281
- Carmo, M. do, 83
- Cartan, É., 224
- Catenoid, 274
- Cauchy Integral Formula, Generalized, 323
- Cauchy Integral Theorem, Generalized, 322
- Central projection, 17

- Chern, S.-S., 82
- Classical Jacobi equation, 212
- Classical vector analysis, 128
- Closed geodesic, 312
- Closed path, 240
 - smooth, 240
 - special, 244
- Codazzi-Mainardi equations, 33, 36, 38, 43
 - generalized, 173
- Comparison theorem
 - Morse-Schoenberg, 234
 - Rauch, 236
 - Sturm, 226
 - delicate, 258
- Complete hypersurfaces
 - flat in \mathbb{R}^{n+1} , 126
 - of constant curvature 1 in S^{n+1} , 126
 - of constant curvature -1 in H^{n+1} , 127
- Complete surfaces
 - of constant curvature in H^3 , 110–119
 - of constant curvature in S^3 , 93–110
 - with $K_{\text{int}} = K_0$, 88
- Complex derivatives, formal, 319
- Conformal model of hyperbolic space, 7–16
- Conjugate
 - locus, 248
 - point, 210
 - geodesic without, 217
 - value, 210
- Connection
 - curvature of, 41
 - normal, 35
- Constant curvature
 - complete hypersurfaces of, 126
 - complete surfaces of, 88, 93–119
- Constant geodesic curvature, 307, 309
- Convex, 81–83
- Courant, R., 281
- Covering transformation, 242
- Cross ratio, 190
- Curvature *see also* Constant curvature
 - elementary symmetric, 65, 71

- Curvature (*continued*)
 extrinsic Gaussian, 86
 function, 21
 Gaussian, 65
 geodesic, constant, 307, 309
 intrinsic, 86
 mean, 65
 of an Einstein space, 121
 normal, mean, 71
 of a connection, 41
 positive, 81 ff.
 principal, 64, 70
 for a normal vector, 71
 Curved, nicely, 166
 Cut
 locus, 248
 point, 248
 $C^{n+\alpha}$, 318
 C^1 functions, Euler's Rule for, 299
 C^α , 318
- Darboux frame, 88
 Deck transformation, 242
 Decomposition Theorem, Hodge, 144
 Definite, *see* Negative definite
 Delicate Sturm comparison theorem, 258
 Derivative, formal complex, 319
 Dido, problem of, 294
 Differential equation, *see* First order
 linear equation, general
 Direct methods in calculus of variations, 241
 Directions
 asymptotic, 89
 principal, 64, 70
 for a normal vector, 71
 Distribution along a curve, 28
 Divergence, 128 ff.
 Divergence Theorem, 130, 132
 Do Carmo, *see* Carmo, M. do
 Douglas, J., 280
 Dual problem, 294
 Dupin's Theorem, 8
- Einstein space, 121
 Eisenhardt, L. P., 90
 Elementary symmetric
 curvatures, 65, 71
 functions, 65
 Endpoints fixed, 205
 Enneper, A., 89, 270
 Enneper's minimal surface, 274
 Envelope of geodesics, 220
 Equidistant hypersurfaces, 16
 Euler's equation, analogue of, 283
 Euler's Rule, 298
 for C^1 functions, 299
 Extrinsic Gaussian curvature, 86
- Fialkow, A., 123
 Films, soap, 280 ff.
 First order linear equation, general, 314
 Fixed endpoints, 205
 Flat
 analytic surfaces in \mathbb{R}^3 , 118
 hypersurfaces in \mathbb{R}^{n+1} , complete, 126
 non-ruled surfaces in \mathbb{R}^m , 86
 ruled surfaces in \mathbb{R}^m , 86
 torus, 106
 Formal
 complex derivative, 319
 imbedding number, 166
 Fractional transformation, linear, 189
 Franke, 52
 Darboux, 88
 Frenet, 24
 Free homotopy class, 241
 Frenet, 22 ff.
 equations, 172
 frame, 24
 Fundamental forms
 normal, 35
 second, 35
 Fundamental Lemma of Riemannian
 Submanifold Theory, 174

- Gauss, C. E., 314
- Gauss formulas, 32
- Gauss' equation, 32, 35
 - generalized, 180
- Gaussian curvature, 65
 - extrinsic, 86
 - formula for, 69
- General first order linear equation, 314
- Generalized
 - Cauchy Integral Formula, 323
 - Cauchy Integral Theorem, 322
 - Codazzi-Mainardi equations, 173
- Generalized Gauss equation, 180
- Geodesic *see also* Totally geodesic
 - closed, 312
 - curvature, constant, 307, 309
 - mapping, 17–19, 192–194
 - of hyperbolic space, 4
 - spheres of hyperbolic space, 11, 13, 16
 - without conjugate points, 217
- Geodesics, envelope of, 220
- Gradient, 128–129
- Green's Theorem, 132
- Gromoll, D., 239

- Hadamard-Cartan, 224
- Half-space model of hyperbolic space, 13
- Harmonic form, 144
- Hartshorne, R., 192
- Helices, 110
- Helicoid, 274
- Henneberg's minimal surface, 278
- Herglotz, G., 313
- Hilbert, D., 241
- Hodge Decomposition Theorem, 144
- Hölder condition, 318
- Homotopy class, free, 241
- Hopf, H., 310, 350
- Hopf map, 107
- Hopf's Problem, 311
- Horosphere, 14, 16

- Hyperbolic space, 2 ff.
 - conformal model of, 7–16
 - equidistant hypersurfaces of, 16
 - geodesic mappings of, 18
 - geodesic spheres of, 11, 13, 16
 - horospheres of, 14, 16
 - isometries of, 4, 10
 - limit spheres of, *see* Horospheres
 - projective model of, 19
 - totally geodesic submanifolds of, 4, 11
 - upper half-space model of, 13

- Imbedding number, formal, 166
- Index
 - of a bilinear function, 3
 - of E_{**} , 223
- Index Theorem, Morse, 223
- Inner product, Lorentzian, 2
- Integral
 - Formula, Generalized Cauchy, 322
 - Theorem, Generalized Cauchy, 323
- Integrating factor, 314
- Intrinsic
 - curvature, 86
 - Riemannian geometry, 128
- Isometries
 - of conformal model of hyperbolic space, 10
 - of hyperbolic space, 4
 - of S^n , 2
 - of upper half-space model of hyperbolic space, 13
- Isomorphism, bundle, 48
- Isoperimetric problem, 294 ff.
- Isothermal coordinate system, 264

- Jacobi equation, 208
 - classical, 212
- Jacobi field, 208
- Jörgens, K., 112

- K , formula for, 69
- Kapouleas, N., 311
- Klingenberg, W., 239, 313
- Klingenberg's Theorem, 255
- Kühne, H., 38

- Lagrange, J. L., 284
- Lagrangian multipliers, 295
- Laplace's equation, 268
- Laplace-Beltrami operator, 133
- Laplacian, 128 ff., 143
- Lashof, R. K., 82
- Lawson, H. B., 293
- Leibniz's Rule, 286
- Lewy (H.), transformation of, 113, 266
- Lie algebra of S^3 , 97
- Lima, E., 83
- Limit sphere, *see* Horosphere
- Linear equation, general first order, 314
- Linear fractional transformation, 189
- Liouville's theorem, 9
- Locus
 - conjugate, 248
 - cut, 248
- Lorentz group, 2
- Lorentzian inner product, 2

- Mainardi, G., *see* Codazzi-Mainardi equations
- Mayer, W., 163
- Mean curvature, 65
 - normal, 71
 - of an Einstein space, 121
- Meusnier, J. B., 285
- Meyer, W., 239
- Milnor, J. W., 223
- Minding, F. A., 307
- Minimal surface
 - associated, 275
 - catenoid, 274
 - Enneper's, 274
 - helicoid, 274
 - Henneberg's, 278
 - Scherk's, 274
- Morse index theorem, 223
- Morse theory, 82
- Morse-Schoenberg Comparison Theorem, 234
- Multiplicity of conjugate values, 210
- Multipliers, Lagrangian, 295
- Myers' Theorem, 235

- Negative definite, 223
- Nicely curved, 166
- Non-ruled flat surfaces in \mathbb{R}^m , 86
- Normal
 - bundle, of submanifold, 33
 - connection, 35
 - fundamental forms, 35
 - mean curvature, 65
 - outward pointing unit, 370
 - space, 167
 - variation, 260

- O'Neil, B., 127
- Orthogonal systems of hypersurfaces in \mathbb{R}^n , 8
- Osculating
 - plane, 24
 - space, 166
- Osserman, R., 310
- Outward pointing unit normal, 370

- Parallel
 - along curve, 28
 - curve, 293
 - surface, 292
- Partial Ricci tensor, 355
- Pinched (δ -pinched), 238
- Plateau, J., 280
- Plateau problem, 280
- Poincaré, H., 312
- Positive curvature, 81 ff.
- Principal
 - bundle isomorphism, 54
 - curvatures, 64, 70
 - for a normal vector, 71
 - directions, 64, 70
 - for a normal vector, 71
- Product tori in S^3 , 109, 110
- Projection
 - central, 17
 - stereographic, 5, 107, 265
- Projective model of hyperbolic space, 19

- Quaternions, 96

- Rado, T., 280, 373; *see also* Schwarz-Rado
- Ratio, cross, 190
- Rauch, H. E., 236
- Rauch Comparison Theorem, 236
- Ricci, G., 377
- Ricci
 - equations, 38, 39, 41, 43
 - tensor, 120
 - partial, 355
- Ricci-Kühne equations, 38
- Riemannian geometry, intrinsic, 128
- Riemannian Submanifold Theory, Fundamental Lemma of, 174
- Ruled surfaces
 - in Riemannian manifolds, 86
 - in \mathbb{R}^m , 85
 - flat, 86
- Ryan, P.J., 123

- Sacksteder, R., 82
- Sasaki, S., 117
- Scherk's minimal surface, 274
- Schoenberg, J.M., 234
- Schwarz, H., 275
- Schwarz-Rado Theorem, 373
- Second
 - Bianchi identity, 182
 - fundamental forms, 35
 - variation formula, 205
 - of volume, 355
- Serret-Frenet formulas, 21 ff.
- Smooth closed path, 240
- Soap
 - bubbles, 310
 - films, 280 ff.
- Special closed path, 244
- Sphere, 1
 - bundle, 251
 - geodesic, in hyperbolic space, 11, 13
- Sphere theorem, 239
- Steiner, J., 302
- Stereographic projection, 5, 107, 265
- Sturm, J. C. F., 230
- Sturm
 - Comparison Theorem, 226
 - delicate, 258
- Symmetric
 - curvatures, elementary, 65, 71
 - functions, elementary, 65
- Synge's Lemma, 239
- Synge's Theorem, 241
- System of hypersurfaces in \mathbb{R}^n , orthogonal, 8
- S^3
 - Lie algebra of, 97
 - product tori in, 109
 - translation in, 96

- Tangent space of Ω at γ , 203
- Theorema Egregium, 66, 70
- Three body problem, 312
- Torus
 - flat, 106
 - product in S^3 , 109
- Totally geodesic submanifolds of
 - hyperbolic space, 4, 11, 13
- Transformation
 - covering (= deck), 242
 - linear fractional, 189
 - of Lewy, 113, 266
- Translation
 - in S^3 , 96
 - surface, 107
- Two-parameter variation, 205

- Umbilic, 8, 72
 - hypersurface of \mathbb{R}^m with all points, 8
 - submanifold of H^m with all points, 77
 - submanifold of \mathbb{R}^m with all points, 75
 - submanifold of S^m with all points, 75
- Unit normal, outward pointing, 370

- Upper half-space model of hyperbolic space, 13

- Variation, 260; *see also* Calculus of variations
 - formula, second, 205
 - normal, 260
 - of volume formula, 287
 - second, 355
 - two-parameter, 205
 - vector field, 205, 260
- Vector analysis, classical, 128
- Vector field, variation, 205, 260
- Vladimirova, S. M., 117
- Volkov, Ju. A., 117
- Volume
 - form, 287
 - variation of, formula for, 287
 - variation of, second, 355

- Warner, F. W., 83, 144
- Weierstrass, K., 270
- Weingarten equations, 33, 35
- Wente, H. C., 311

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME FIVE
Third Edition



MICHAEL SPIVAK

PUBLISH OR PERISH, INC.



Houston, Texas 1999

Publish or Perish, Inc.
www.mathpop.com

Copyright © 1970, 1979, 1999 by Michael Spivak
All Rights Reserved

Volume 1 ISBN 0-914098-70-5
Volume 2 ISBN 0-914098-71-3
Volume 3 ISBN 0-914098-72-1
Volume 4 ISBN 0-914098-73-X
Volume 5 ISBN 0-914098-74-8

Printed in the United States of America

TABLE OF CONTENTS

Although the chapters are not divided into sections,
except for the subdivisions of Chapters 10 and 13,
the listing for each chapter gives some indication
which topics are treated, and on what pages.

CHAPTER 10. AND NOW A BRIEF MESSAGE FROM OUR SPONSOR

1. FIRST ORDER PDE's	1
Linear first order PDE's; characteristic curves; Cauchy problem	
for free initial curves	4
Quasi-linear first order PDE's; characteristic curves; Cauchy problem	
for free initial conditions; characteristic initial conditions	9
General first order PDE's; Monge cone; characteristic curves	
of a solution; characteristic strips; Cauchy problem for	
free initial data; characteristic initial data	13
First order PDE's in n variables	26
2. FREE INITIAL MANIFOLDS FOR HIGHER ORDER EQUATIONS	29
3. SYSTEMS OF FIRST ORDER PDE's	36
4. THE CAUCHY-KOWALEWSKI THEOREM	38
5. CLASSIFICATION OF SECOND ORDER PDE's	
Classification of semi-linear equations	47
Reduction to normal forms	50
Classification of general second order equations	56
6. THE PROTOTYPICAL PDE's OF PHYSICS	
The wave equation; the heat equation; Laplace's equation	59
Elementary properties	66
7. HYPERBOLIC SYSTEMS IN TWO VARIABLES	72
8. HYPERBOLIC SECOND ORDER EQUATIONS IN TWO VARIABLES	
First reduction of the problem	81
New system of characteristic equations	84

Characteristic initial data	96
Monge-Ampère equations	97
9. ELLIPTIC SOLUTIONS OF SECOND ORDER EQUATIONS	
IN TWO VARIABLES	98
Addendum 1. Differential systems; the Cartan-Kähler Theorem . . .	110
Addendum 2. An elementary maximum principal	126
Problem	132
CHAPTER 11. EXISTENCE AND NON-EXISTENCE OF ISOMETRIC IMBEDDINGS	
Non-imbeddability theorems; exteriorly orthogonal bilinear forms; index of nullity and index of relative nullity	133
The Darboux equation	142
Burstin-Janet-Cartan Theorem	147
Addendum. The embedding problem via differential systems . . .	157
Problems	165
CHAPTER 12. RIGIDITY	
Rigidity in higher dimensions; type number	167
Bendings, warpings, and infinitesimal bendings	170
\mathbb{R}^3 -valued differential forms, the support function, and Minkowski's formulas	181
Infinitesimal rigidity of convex surfaces	186
Cohn-Vossen's Theorem	192
Minkowski's Theorem	200
Christoffel's Theorem	204
Other problems, solved and unsolved	206
Local problems; the role of the asymptotic curves	215
Other classical results	221
E. E. Levi's Theorems and Schilt's Theorem	227
Surfaces in S^3 and H^3	235
Rigidity for higher codimension	247
Addendum. Infinitesimal bendings of rotation surfaces	253
Problems	261

CHAPTER 13. THE GENERALIZED GAUSS-BONNET THEOREM AND WHAT IT MEANS FOR MANKIND

Historical remarks	263
1. OPERATIONS ON BUNDLES	
Bundle maps and principal bundle maps; Whitney sums and induced bundles; the covering homotopy theorem	265
2. GRASSMANNIANS AND UNIVERSAL BUNDLES	273
3. THE PFAFFIAN	284
4. DEFINING THE EULER CLASS IN TERMS OF A CONNECTION	
The Euler class	291
The class $C(\xi)$	294
The Gauss-Bonnet-Chern Theorem	297
5. THE CONCEPT OF CHARACTERISTIC CLASSES	302
6. THE COHOMOLOGY OF HOMOGENEOUS SPACES	
The C^∞ structure of homogeneous spaces	304
Invariant forms	308
7. A SMATTERING OF CLASSICAL INVARIANT THEORY	
The Capelli identities	317
The first fundamental theorem of invariant theory for $O(n)$ and $SO(n)$	326
8. AN EASIER INVARIANCE PROBLEM	330
9. THE COHOMOLOGY OF THE ORIENTED GRASSMANNIANS	
Computation of the cohomology; Pontryagin classes	337
Describing the characteristic classes in terms of a connection	345
10. THE WEIL HOMOMORPHISM	353
11. COMPLEX BUNDLES	
Hermitian inner products, the unitary group, and complex Grassmannians	356
The cohomology of the complex Grassmannians; Chern classes	360
Relations between the Chern classes and the Pontryagin and Euler classes	365
12. VALEDICTORY	369
Addendum 1. Invariant theory for the unitary group	372
Addendum 2. Recovering the differential forms; the Gauss-Bonnet-Chern Theorem for manifolds-with-boundary	380

BIBLIOGRAPHY	393
A. Other topics in Differential Geometry	394
B. Books	411
C. Journal articles	445
NOTATION INDEX	459
INDEX	461

A
Comprehensive Introduction
to
DIFFERENTIAL GEOMETRY

VOLUME FIVE

CHAPTER 10

AND NOW A BRIEF MESSAGE FROM OUR SPONSOR

Partial differential equations have played a decisive role in our investigations ever since they were first introduced in Chapter 6 of Volume I. To be sure, at times we have suppressed the equations themselves in favor of a more geometric conception involving k -dimensional distributions, and on other occasions we have instead expressed things in terms of differential forms. But, in one form or another, the Frobenius Theorem (which represents everything we know about partial differential equations) was used in discussing Lie groups, ordinary and affine theory of curves and surfaces in space (where Lie group methods were used), in all our proofs of the Test Case, in the proof of the Fundamental Theorem of Surface Theory, and in the generalizations of this theorem which were given in Chapter 7. The partial differential equations involved are of the form

$$\frac{\partial \alpha^i}{\partial x_j}(x_1, \dots, x_m) = f_j^i(x_1, \dots, x_m, \alpha^1(x_1, \dots, x_m), \dots, \alpha^n(x_1, \dots, x_m))$$

$$i = 1, \dots, n; \quad j = 1, \dots, m.$$

Now it's really rather laughable to call these things partial differential equations at all. True, we are considering functions α^i defined on \mathbb{R}^n , and therefore partial derivatives are involved, but the equations do not posit any relationship between *different* partial derivatives; this comes out quite clearly in the proof, where the equations are reduced to ordinary differential equations. The only reason we get anything interesting at all in this situation is because we are dealing with a *system* of equations, and this system is "overdetermined": there are more equations (namely mn) than there are unknown functions (namely n). Our particular overdetermined system happens to be one where it is not too hard to determine the additional "integrability conditions" which must hold for the functions f_j^i if the strain of satisfying so many equations is not to hopelessly overburden the poor functions α^i .

With only this superficial knowledge of partial differential equations, one can make one's way through a good part of differential geometry ("*the good part*", you may be inclined to say after looking at this chapter). But there are some topics in differential geometry, to be covered in the next two chapters, where

a more intimate acquaintance with partial differential equations is required. It should be said right away, that even in the next two chapters there are only a few occasions where this knowledge is necessary, and one could easily decide to take on faith any theorems from this chapter which happen to be quoted later. On the other hand, many theorems cannot even be stated without some definitions that arise in the first attempts to understand partial differential equations; these definitions involve basic facts about the behavior of partial differential equations, and this behavior is often reflected in geometric phenomena in a surprisingly nice way.

This chapter is not meant to be a substitute for a course in partial differential equations; we will try to reach in as short a space as possible those particular properties of partial differential equations which will be of importance to us in the next two chapters, even if they are of only secondary importance to analysis. Consequently, we will omit much material that is contained in elementary courses, and at the same time prove special cases of results which are usually found only in more advanced treatments, where they are proved in much greater generality, and with much more effort. (Just to keep the presentation from being too one-sided, passing mention has sometimes been given to matters which are of great importance to analysts, but of no importance to us). Since we are going to be totally immersed in the study of partial differential equations for quite a while, we might as well admit it, and henceforth resort to the standard abbreviation PDE.

A few general considerations might be made before we begin in earnest. When we consider an ordinary differential equation

$$u'(x) = f(x, u(x)),$$

we find that there are solutions u with any desired value for $u(x_0)$. This dependence on the “initial condition” $u(x_0)$ usually manifests itself, if we explicitly solve the equation, by the presence of an arbitrary constant of integration. For example, the equation

$$\frac{du}{dx} = -u^2 \quad \left(\Rightarrow \frac{du}{-u^2} = dx \Rightarrow \frac{1}{u} = x + C \right)$$

has the “general” solution

$$u(x) = \frac{1}{x + C},$$

which gives all desired initial conditions $u(x_0)$ except $u(x_0) = 0$; for this one needs the “singular” solution $u(x) = 0$. Equations of order n , on the other hand, will involve n constants of integration.

When we solve a PDE, we usually obtain arbitrary *functions* in the answer. For example, to be as simple-minded about the thing as we can, we note that the equation

$$\frac{\partial u}{\partial y}(x, y) = 0$$

has the solutions $u(x, y) = A(x)$; the only restrictions on A are ones which follow from restrictions we might choose to place on u (e.g., that u be differentiable with respect to x). The equally stupid looking, but actually quite important, second order equation

$$\frac{\partial^2 u}{\partial x \partial y}(x, y) = 0$$

leads to

$$\frac{\partial u}{\partial x}(x, y) = \alpha(x),$$

and hence to

$$u(x, y) = A(x) + B(y), \quad A'(x) = \alpha(x).$$

Without belaboring the point any further, we simply note that when we look for precise theorems, we should expect the hypotheses to reflect the presence of these “arbitrary functions” in the same way that the precise theorem for ordinary differential equations reflects the presence of arbitrary constants.

1. FIRST ORDER PDE's

In this section we will consider those equations which involve a function u on \mathbb{R}^n and only its first partial derivatives u_{x_i} . For simplicity of writing, and convenience of visualization, we will first deal exclusively with the case of \mathbb{R}^2 , denoting a typical point of \mathbb{R}^2 by (x, y) and adopting the standard notation

$$u_x = p, \quad u_y = q.$$

By a **first order PDE** we then mean an equation of the form

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0,$$

or, to use the standard abbreviated form,

$$F(x, y, u, p, q) = 0.$$

It will be convenient to denote the various partial derivatives of F by F_x , F_y , F_u , F_p , and F_q . Naturally, the function $F: \mathbb{R}^5 \rightarrow \mathbb{R}$ shouldn't be too badly

behaved; for example, it wouldn't be very interesting if F were never 0. Just what hypotheses we really need will come out soon enough. To begin with, we might imagine that F is differentiable and satisfies $F_p \neq 0$ or $F_q \neq 0$, so that by the implicit function theorem we can solve for p in terms of q , or *vice versa*. Our main result is, that we can always completely reduce any first order PDE to a system of ordinary differential equations. This holds both in a "practical" and in a theoretical sense: We can actually write down a system of ordinary differential equations whose solutions, if we can find them, will give us the solution of our original problem; and the method by which this is done enables us to state and prove exact theorems. We will not deal at the very outset with the most general first order PDE, but will approach it in stages.

We consider first the most general **linear first order PDE**

$$(1) \quad A(x, y)u_x(x, y) + B(x, y)u_y(x, y) = C(x, y)u(x, y) + D(x, y).$$

Usually this is simply written

$$A(x, y)u_x + B(x, y)u_y = C(x, y)u + D(x, y),$$

with the arguments (x, y) appearing in A , B , C , and D just to emphasize that we are not considering an equation like $A(x, y, u(x, y))u_x + \dots$.

Consider the vector field X on \mathbb{R}^2 defined by

$$(2) \quad X = A \frac{\partial}{\partial x} + B \frac{\partial}{\partial y}.$$

The value of X at (x_0, y_0) is

$$A(x_0, y_0) \frac{\partial}{\partial x} \Big|_{(x_0, y_0)} + B(x_0, y_0) \frac{\partial}{\partial y} \Big|_{(x_0, y_0)};$$

using the standard identification of the tangent space $\mathbb{R}^2_{(x_0, y_0)}$ with \mathbb{R}^2 , we can also write

$$X(x_0, y_0) = (A(x_0, y_0), B(x_0, y_0)).$$

We will call X the **characteristic vector field** of equation (1); the integral curves of this vector field are called the **characteristic curves** of equation (1). Thus $c = (c_1, c_2)$ is a characteristic curve if and only if

$$(3) \quad \frac{dc_1(t)}{dt} = A(c(t)), \quad \frac{dc_2(t)}{dt} = B(c(t)).$$

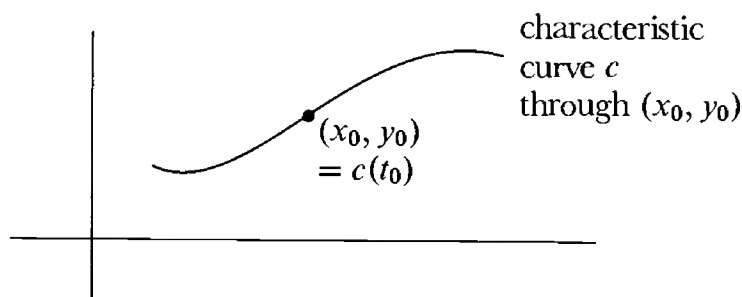
We then have, for any C^1 function $u: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\begin{aligned} \frac{du(c(t))}{dt} &= u_x(c(t)) \frac{dc_1(t)}{dt} + u_y(c(t)) \frac{dc_2(t)}{dt} \\ &= A(c(t)) \cdot u_x(c(t)) + B(c(t)) \cdot u_y(c(t)). \end{aligned}$$

So any solution u of equation (1) satisfies

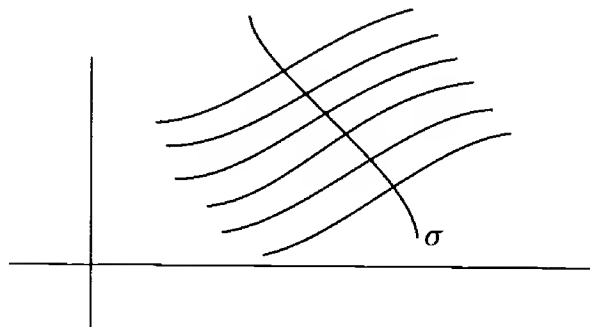
$$(4) \quad \frac{du(c(t))}{dt} = C(c(t)) \cdot u(c(t)) + D(c(t)) \quad \text{for any characteristic curve } c.$$

For any fixed characteristic curve $t \mapsto c(t)$, equation (4) is an ordinary differential equation for the function $u \circ c$. Consequently, $u \circ c$ is uniquely determined



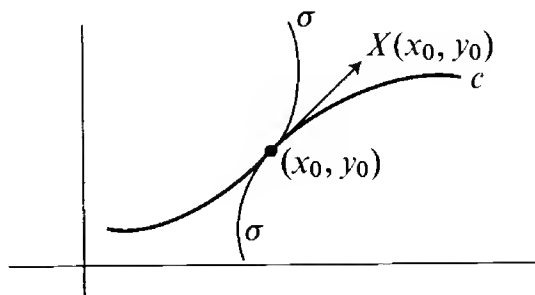
once $u(c(t_0))$ is specified. In other words, once we prescribe a value $u(x_0, y_0)$ for a solution u of equation (1), the solution u will then be completely determined along the characteristic curve c through (x_0, y_0) .

Now suppose we have any curve σ which cuts a family of characteristic curves.



If we arbitrarily specify the values of u at each point of σ , then the solution u will be determined in a neighborhood of σ . Moreover, we ought to be able to produce this solution u simply by solving equation (4) for each of the characteristic curves through each point of σ . Of course, we clearly have to rule out the possibility that a portion of σ itself is a characteristic curve, for then

we could not arbitrarily specify the values of u along σ . We even have to rule out the possibility that σ is tangent to some integral curve c at some point $(x_0, y_0) = c(t_0)$; for in this case, the directional derivative $X(x_0, y_0)(u)$ would



be determined both by equation (4) and (in a possibly conflicting way) by the arbitrarily assigned values of u along σ . We must thus assume that the vectors

$$\sigma'(s) = (\sigma_1'(s), \sigma_2'(s)) \quad \text{and} \quad (A(\sigma(s)), B(\sigma(s)))$$

are always linearly independent. Equivalently, we must require that

$$0 \neq \det \begin{pmatrix} \sigma_1'(s) & A(\sigma(s)) \\ \sigma_2'(s) & B(\sigma(s)) \end{pmatrix} = \sigma_1'(s)B(\sigma(s)) - \sigma_2'(s)A(\sigma(s))$$

for all s . In particular, $\sigma'(s) \neq (0, 0)$ so σ is an imbedding. Although we will later have a much more general result, we summarize this information in a theorem, in order to get all the details cleaned up before we carry the discussion any further.

1. THEOREM. Let A , B , C , and D be C^k functions defined in an open set $U \subset \mathbb{R}^2$, and let $\sigma: [a, b] \rightarrow U$ be a one-one C^k curve such that

$$\sigma_1'(s)B(\sigma(s)) \neq \sigma_2'(s)A(\sigma(s)) \quad \text{for all } s \in [a, b].$$

Let $\mathring{u}: [a, b] \rightarrow \mathbb{R}$ be a C^k function. Then there is a C^k function u , defined in a neighborhood V of $\sigma([a, b])$, such that u satisfies

$$(1) \quad A \cdot u_x + B \cdot u_y = C \cdot u + D \quad \text{on } V,$$

with the initial condition

$$u(\sigma(s)) = \mathring{u}(s) \quad \text{for all } s \in [a, b].$$

Moreover, any two functions u with this property agree on a neighborhood of $\sigma([a, b])$.

PROOF. There is a C^k map

$$\gamma : [a, b] \times (-\varepsilon, \varepsilon) \rightarrow U$$

such that each curve

$$t \mapsto \gamma(s, t)$$

is a characteristic curve with

$$\gamma(s, 0) = \sigma(s).$$

Clearly

$$\frac{\partial \gamma}{\partial s}(s, 0) = \sigma'(s) = (\sigma_1'(s), \sigma_2'(s))$$

$$\frac{\partial \gamma}{\partial t}(s, 0) = (A(\sigma(s)), B(\sigma(s))).$$

So, by the hypothesis on σ , the Jacobian of γ at $(s, 0)$ is always non-singular; consequently, if ε is sufficiently small, then γ is a C^k diffeomorphism onto a neighborhood V of $\sigma([a, b])$.

By choosing ε still smaller, if necessary, we can insure that for each $s \in [a, b]$ there is a C^k function $\beta_s : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$ satisfying

$$\begin{cases} \frac{d\beta_s(t)}{dt} = C(\gamma(s, t)) \cdot \beta_s(t) + D(\gamma(s, t)) \\ \beta_s(0) = \dot{u}(s) \end{cases}$$

[this is just the equation (4) which should be satisfied by $u \circ c$ along the integral curve $t \mapsto \gamma(s, t)$]. We would actually like to know that $\beta_s(t)$ is C^k as a function of s and t ; in other words, if we define $\beta : [a, b] \times (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$ by

$$\beta(s, t) = \beta_s(t),$$

then we would like to know that β is C^k . To prove this, we must consider the equation “depending on parameters”

$$\begin{cases} \alpha(0, s, r) = r & \text{for } r \in \mathbb{R} \\ \frac{\partial}{\partial t} \alpha(t, s, r) = C(\gamma(s, t)) \cdot \alpha(t, s, r) + D(\gamma(s, t)). \end{cases}$$

Problem I.5-5 shows that α is C^k ; consequently

$$\beta(s, t) = \alpha(t, s, \dot{u}(s))$$

is also C^k .

Now the solution u , if it exists, clearly must be the C^k function

$$u(x, y) = \beta(\gamma^{-1}(x, y)) \quad \text{or equivalently} \quad u(\gamma(s, t)) = \beta(s, t).$$

To prove that u really is a solution, we note that through any point $(x, y) \in V$ there is a characteristic curve $t \mapsto \gamma(s, t)$, and that

$$\begin{aligned} \frac{du(\gamma(s, t))}{dt} &= \frac{d\beta(s, t)}{dt} = C(\gamma(s, t)) \cdot \beta(s, t) + D(\gamma(s, t)) \\ &= C(\gamma(s, t)) \cdot u(\gamma(s, t)) + D(\gamma(s, t)), \end{aligned}$$

while we also have

$$\begin{aligned} \frac{du(\gamma(s, t))}{dt} &= u_x(\gamma(s, t)) \cdot \frac{\partial \gamma_1}{\partial t}(s, t) + u_y(\gamma(s, t)) \cdot \frac{\partial \gamma_2}{\partial t}(s, t) \\ &= u_x(\gamma(s, t)) \cdot A(\gamma(s, t)) + u_y(\gamma(s, t)) \cdot B(\gamma(s, t)), \end{aligned}$$

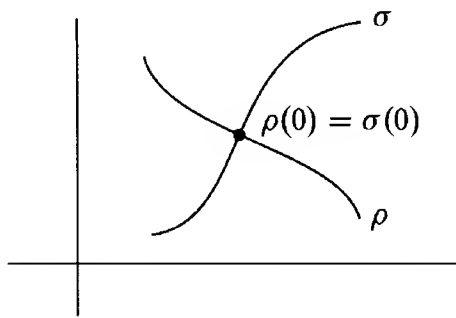
since $t \mapsto \gamma(s, t)$ is a characteristic curve. ♦

Notice that Theorem 1 involves exactly the sort of “arbitrary function” that our general considerations would lead us to expect: in a neighborhood of the “initial curve” σ , the solution u is uniquely determined by the “initial condition” $u(\sigma(s)) = \hat{u}(s)$. The only requirement is that σ be nowhere tangent to a characteristic curve; we will express this by saying that σ is **free** (sometimes the term “non-characteristic” is used, but this seems a little misleading). In general, the problem of finding a solution of a PDE with an appropriate initial condition is called the “Cauchy problem” for this equation. Thus we have solved the Cauchy problem for the linear PDE (1) for any initial condition along any free curve. In particular, we can solve the Cauchy problem along the x -axis $\sigma(s) = (s, 0)$ if the x -axis is free, which is equivalent to the condition that $B \neq 0$ along the x -axis. In this case we can use the given equation (1) to solve for u_y in terms of u_x along the x -axis:

$$u_y = -\frac{A}{B}u_x + \frac{C}{B}u + \frac{D}{B}.$$

If we were interested in the Cauchy problem only along the x -axis, then we could simply demand this very natural condition in our hypotheses, and not mention the characteristic curves at all; but the characteristic curves are still the most important ingredient in the proof, and their generalizations will play decisive roles in all other equations we discuss.

If our initial curve σ actually happens to be a characteristic curve (thus failing in the worst possible way to be free), then we will be unable to solve the Cauchy problem, and this inability will be manifested in the worst possible way: the possible initial condition along σ is almost uniquely determined—it is determined by the value at only one point, by the equation (4). On the other hand, if we are given an initial condition \bar{u} along σ which does satisfy (4), then there will be infinitely many solutions u with this initial condition; for we can consider any free curve ρ with $\rho(0) = \sigma(0)$, and choose any initial data ϕ along ρ



with $\phi(0) = \bar{u}(0)$. Thus, the characteristic curves are the places where different solutions agree.

From Theorem 1 we can see immediately that an arbitrary linear first order PDE has, in common with the simple-minded equation $\partial u / \partial y = 0$, a property which sharply distinguishes it from an *ordinary* differential equation

$$u'(x) = f(x, u(x)).$$

For the ordinary differential equation, any solution u will clearly be at least one time more differentiable than f is, and if f is analytic, the solution will also be analytic (Problem I.6-9). But there are solutions of the equation in Theorem 1 which are only C^l ($1 \leq l \leq \infty$) even when A, B, C, D are C^k ($l < k \leq \omega$). For we may choose σ to be a C^k curve and \bar{u} to be a function which is C^l , but not C^{l+1} ; then the solution u cannot be C^{l+1} , since its restriction to the C^k curve σ is not C^{l+1} .

We next consider the most general **quasi-linear first order PDE**

$$A(x, y, u(x, y))u_x(x, y) + B(x, y, u(x, y))u_y(x, y) = C(x, y, u(x, y)),$$

or, more briefly,

$$A(x, y, u)u_x + B(x, y, u)u_y = C(x, y, u).$$

The functions A , B , and C are now defined on \mathbb{R}^3 , and we consider the vector field X in \mathbb{R}^3 defined by

$$(2) \quad X = A \frac{\partial}{\partial x} + B \frac{\partial}{\partial y} + C \frac{\partial}{\partial z}.$$

This vector field will be called the **characteristic vector field** of equation (1); the integral curves of X are called the **characteristic curves** of equation (1). Thus $c = (c_1, c_2, c_3)$ is a characteristic curve if and only if

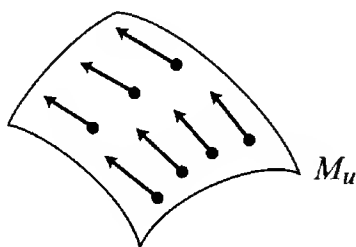
$$(3) \quad \frac{dc_1(t)}{dt} = A(c(t)), \quad \frac{dc_2(t)}{dt} = B(c(t)), \quad \frac{dc_3(t)}{dt} = C(c(t)).$$

The slight discrepancy between this terminology and that adopted in the linear case is easily explained. Notice that if A and B depend only on x and y , then all characteristic vectors $X(x_0, y_0, z_0)$ have the same projection on the (x, y) -plane, namely $(A(x_0, y_0), B(x_0, y_0))$. So the characteristic curves of a linear equation are really the projections on the (x, y) -plane of the characteristic curves in \mathbb{R}^3 .

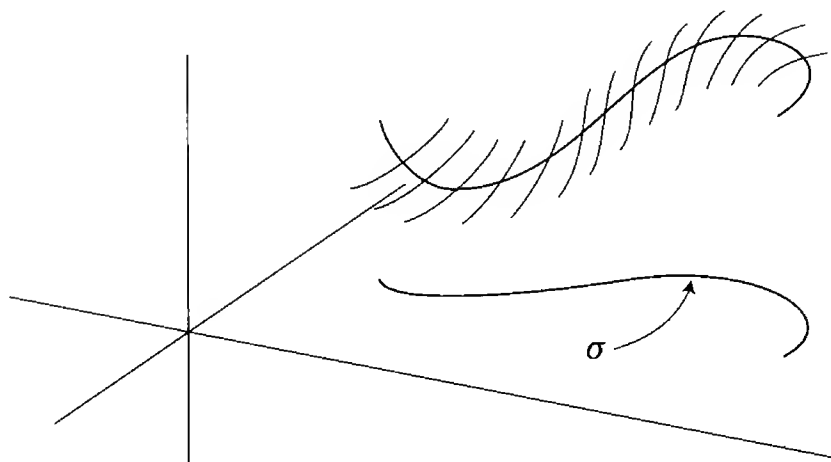
For the quasi-linear PDE (1), the characteristic curves in \mathbb{R}^3 have the following significance. Any C^1 function $u: \mathbb{R}^2 \rightarrow \mathbb{R}$ determines a surface $M_u = \{(x, y, u(x, y))\} \subset \mathbb{R}^3$, and the vector

$$(u_x(x, y), u_y(x, y), -1)$$

is normal to M_u at $(x, y, u(x, y))$. Equation (1) is therefore equivalent to saying that $X(x, y, u(x, y))$ lies in the tangent space of M_u at $(x, y, u(x, y))$. So the



characteristic vectors at the various points of M_u give a vector field on M_u . Thus M_u is the union of integral curves of this vector field; that is, M_u is the union of characteristic curves. If we are given an arbitrary initial condition \hat{u} along an initial curve σ in \mathbb{R}^2 , then we ought to be able to construct a solution u passing through the curve $s \mapsto (\sigma_1(s), \sigma_2(s), \hat{u}(s))$ in \mathbb{R}^3 simply by taking the union of the characteristic curves through all the points of this curve. We will clearly have to require that the vectors $(\sigma_1'(s), \sigma_2'(s))$ and $(A(\sigma_1(s), \sigma_2(s), \hat{u}(s)), B(\sigma_1(s), \sigma_2(s), \hat{u}(s)))$ are linearly independent for all s .



2. THEOREM. Let A , B , and C be C^k functions defined in an open set $U \subset \mathbb{R}^3$. Let $\sigma: [a, b] \rightarrow \mathbb{R}^2$ be a one-one C^k function, and $\dot{u}: [a, b] \rightarrow \mathbb{R}$ a C^k function such that $(\sigma_1(s), \sigma_2(s), \dot{u}(s)) \in U$ for all $s \in [a, b]$. Suppose moreover that

$$\sigma_1'(s) \cdot B(\sigma_1(s), \sigma_2(s), \dot{u}(s)) \neq \sigma_2'(s) \cdot A(\sigma_1(s), \sigma_2(s), \dot{u}(s)) \quad \text{for all } s \in [a, b].$$

Then there is a C^k function u , defined in a neighborhood V of $\sigma([a, b])$, which satisfies the equation

$$(I) \quad A(x, y, u)u_x + B(x, y, u)u_y = C(x, y, u) \quad \text{on } V,$$

with the initial condition

$$u(\sigma(s)) = \dot{u}(s) \quad \text{for all } s \in [a, b].$$

Moreover, any two functions u with this property agree on a neighborhood of $\sigma([a, b])$.

PROOF. By Problem I.5-5 there is a C^k function $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ with

$$(*) \quad \begin{cases} \alpha(0, s, r) = r & \text{for } r \in \mathbb{R}^3 \\ \frac{\partial}{\partial t} \alpha_1(t, s, r) = A(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_2(t, s, r) = B(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_3(t, s, r) = C(\alpha(t, s, r)). \end{cases}$$

Let

$$\beta(s, t) = \alpha(t, s, \sigma_1(s), \sigma_2(s), \dot{u}(s)),$$

so that β is also C^k . In particular,

$$\begin{aligned}\beta(s, 0) &= (\sigma_1(s), \sigma_2(s), \overset{\circ}{u}(s)) \\ &= \bullet, \quad \text{for short}\end{aligned}$$

[so for each s , the curve $t \mapsto \beta(s, t)$ is a characteristic curve through \bullet]. If we define

$$\gamma(s, t) = (\beta_1(s, t), \beta_2(s, t)) \in \mathbb{R}^2,$$

then the Jacobian of γ at $(s, 0)$ is

$$\begin{aligned}\begin{pmatrix} \frac{\partial \beta_1}{\partial s}(s, 0) & \frac{\partial \beta_1}{\partial t}(s, 0) \\ \frac{\partial \beta_2}{\partial s}(s, 0) & \frac{\partial \beta_2}{\partial t}(s, 0) \end{pmatrix} &= \begin{pmatrix} \sigma_1'(s) & \frac{\partial \alpha_1}{\partial t}(0, s, \bullet) \\ \sigma_2'(s) & \frac{\partial \alpha_2}{\partial t}(0, s, \bullet) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1'(s) & A(\bullet) \\ \sigma_2'(s) & B(\bullet) \end{pmatrix} \quad \text{by } (*),\end{aligned}$$

and this is non-singular, by hypothesis. So if ε is sufficiently small, then $\gamma: [a, b] \times (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^2$ is a C^k diffeomorphism onto a neighborhood V of $\sigma([a, b])$.

The solution u , if it exists, clearly must be the C^k function

$$u(x, y) = \beta_3(\gamma^{-1}(x, y)) \quad \text{or equivalently} \quad u(\gamma(s, t)) = \beta_3(s, t).$$

To prove that u is a solution, we note that for any point $(x, y) \in V$, there is a characteristic curve $t \mapsto \beta(s, t)$ through $(x, y, u(x, y))$, and that

$$\frac{du(\gamma(s, t))}{dt} = \frac{d\beta_3(s, t)}{dt} = C(\beta(s, t)) \quad \text{by } (*),$$

while we also have

$$\begin{aligned}\frac{du(\gamma(s, t))}{dt} &= u_x(\gamma(s, t)) \cdot \frac{\partial \gamma_1}{\partial t}(s, t) + u_y(\gamma(s, t)) \cdot \frac{\partial \gamma_2}{\partial t}(s, t) \\ &= u_x(\gamma(s, t)) \cdot \frac{\partial \beta_1}{\partial t}(s, t) + u_y(\gamma(s, t)) \cdot \frac{\partial \beta_2}{\partial t}(s, t) \\ &\quad \text{by definition of } \gamma \\ &= u_x(\gamma(s, t)) \cdot A(\beta(s, t)) + u_y(\gamma(s, t)) \cdot B(\beta(s, t)) \quad \text{by } (*). \quad \spadesuit\end{aligned}$$

We will say that the initial curve σ is **free for the initial condition** $\overset{\circ}{u}$ when it satisfies

$$\sigma_1'(s) \cdot B(\sigma_1(s), \sigma_2(s), \overset{\circ}{u}(s)) \neq \sigma_2'(s) \cdot A(\sigma_1(s), \sigma_2(s), \overset{\circ}{u}(s)).$$

Thus we can solve the Cauchy problem for a quasi-linear PDE (1) for any initial condition along any curve which is free for this initial condition. (In the linear case things are simpler, since the condition that σ be free doesn't depend on the initial condition \mathring{u} .)

The worst way in which the initial curve $\sigma : [a, b] \rightarrow \mathbb{R}^2$ can fail to be free for the initial condition \mathring{u} is when $\sigma'(s) = (\sigma_1'(s), \sigma_2'(s))$ and $(A(\sigma_1(s), \sigma_2(s), \mathring{u}(s)), B(\sigma_1(s), \sigma_2(s), \mathring{u}(s))) = (A(\bullet), B(\bullet))$ are everywhere linearly dependent. In this case, it is customary to say that σ is **characteristic** for \mathring{u} ; this does *not* mean that σ is a characteristic curve (indeed, σ isn't even a curve in \mathbb{R}^3). If we assume that σ is an imbedding, then σ is characteristic if and only if $(A(\bullet), B(\bullet))$ is always a multiple of the tangent vector $\sigma'(s)$; by reparameterizing σ we can then arrange that

$$(A(\bullet), B(\bullet)) = \sigma'(s).$$

Then if \mathring{u} is to be the initial condition for a solution u of (1) we must have

$$\begin{aligned} C(\bullet) &= \sigma_1'(s) \cdot u_x(\sigma(s)) + \sigma_2'(s) \cdot u_y(\sigma(s)) \\ &= \frac{d}{ds} u(\sigma(s)) = \frac{d}{ds} \mathring{u}(s). \end{aligned}$$

These equations show that the reparameterized curve $s \mapsto (\sigma_1(s), \sigma_2(s), \mathring{u}(s))$ must be a characteristic curve; equivalently, the original curve $s \mapsto (\sigma_1(s), \sigma_2(s), \mathring{u}(s))$ must be a characteristic curve up to reparameterization in order for the Cauchy problem to be solvable when σ is characteristic for \mathring{u} . If our initial condition \mathring{u} does have this property, then there will be infinitely many solutions u with this initial condition along σ . The characteristic curves in \mathbb{R}^3 are the places where the graphs of different solutions intersect; the projections of the characteristic curves onto \mathbb{R}^2 are the places where different solutions agree.

It should be clear once again that a quasi-linear first order PDE has solutions which are less differentiable than its coefficients.

We are now ready to consider the most general *first order* PDE

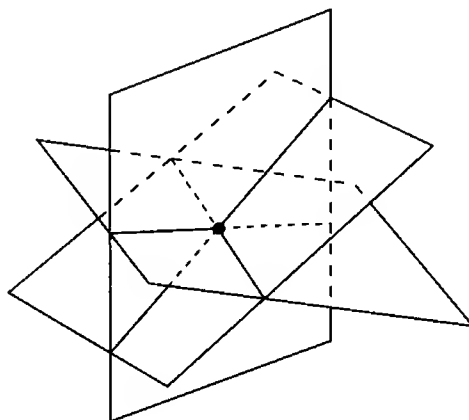
$$(1) \quad F(x, y, u, p, q) = F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0.$$

This equation can also be reduced to a system of ordinary differential equations, but in this case the system will involve *five* functions; the geometric analysis will be correspondingly more complicated.

At each point $(x_0, y_0, z_0) \in \mathbb{R}^3$, we can consider the set of all vectors $(a, b, -1)$ with

$$F(x_0, y_0, z_0, a, b) = 0,$$

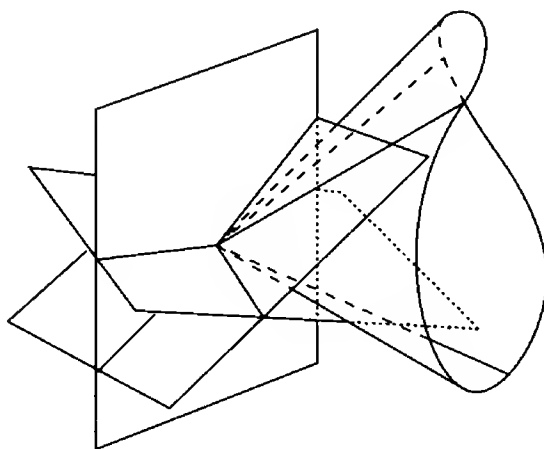
and the corresponding family $\mathcal{F}(x_0, y_0, z_0)$ of planes perpendicular to such vectors. If u is a solution of (1), and M_u is the surface $M_u = \{(x, y, u(x, y))\}$,



then the tangent space of M_u at $(x_0, y_0, u(x_0, y_0))$ is a member of the family $\mathcal{F}(x_0, y_0, u(x_0, y_0))$. In order to describe this situation more geometrically, we would like to have a more geometric way of describing the families $\mathcal{F}(x_0, y_0, z_0)$. Now the relation

$$F(x_0, y_0, z_0, a, b) = 0$$

is one equation in the two unknowns, a and b , so $\mathcal{F}(x_0, y_0, z_0)$ ought to be a one-parameter family of planes; this suggests that there is a cone $K(x_0, y_0, z_0)$, having its vertex at (x_0, y_0, z_0) , with the property that a plane P is in $\mathcal{F}(x_0, y_0, z_0)$ if and only if P is tangent to $K(x_0, y_0, z_0)$ along a generator of this cone. If we



consider a quasi-linear equation

$$F(x, y, u, p, q) = A(x, y, u) \cdot p + B(x, y, u) \cdot q - C(x, y, u) = 0,$$

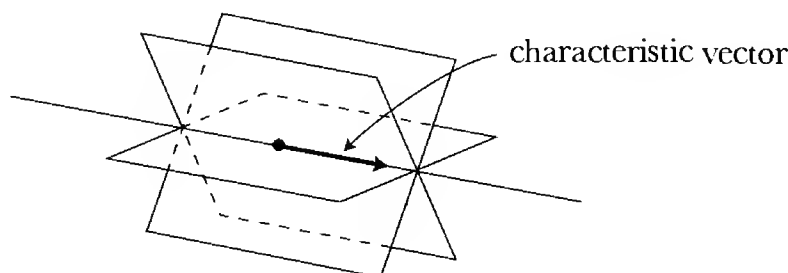
we immediately see that this is not always so. For in this case, the family $\mathcal{F}(x_0, y_0, z_0)$ consists of planes perpendicular to vectors $(a, b, -1)$ with

$$a \cdot A(x_0, y_0, z_0) + b \cdot B(x_0, y_0, z_0) = C(x_0, y_0, z_0).$$

These planes all contain the characteristic vector

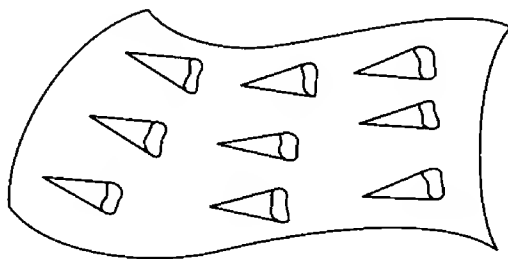
$$(A(x_0, y_0, z_0), B(x_0, y_0, z_0), C(x_0, y_0, z_0)).$$

Thus our “cone” degenerates into a straight line through (x_0, y_0, z_0) , pointing in the direction of the characteristic vector at that point. Clearly things might



be even messier if the analytic properties of the function F are sufficiently nasty.

Despite these difficulties, we can obtain a great deal of geometric motivation by temporarily pretending that each family $\mathcal{F}(x_0, y_0, z_0)$ is determined by a cone $K(x_0, y_0, z_0)$, which happens to degenerate to a straight line in the case of a quasi-linear equation. This semi-mythical cone is called the **Monge cone** at (x_0, y_0, z_0) . Having accepted this fiction, we can now imagine a field of cones in \mathbb{R}^3 ; a C^1 function $u: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a solution of equation (1) if and only



if the corresponding surface $M_u = \{(x, y, u(x, y))\}$ is tangent to the Monge cone $K(x_0, y_0, u(x_0, y_0))$ at each point $(x_0, y_0, u(x_0, y_0))$. This gives us a field of directions at each point of M_u , namely the direction which lies along a generator of the Monge cone at that point. The integral manifolds of this field of directions could be called the “characteristic curves of the solution u ”. This definition is easily seen to be compatible with the one already given in

the quasi-linear case, where the Monge cones degenerate to straight lines. For, these straight lines must be the field of directions for any solution u , and the “characteristic curves of the solution u ” are simply those characteristic curves of the quasi-linear equation which happen to lie on M_u . But in the general case, we cannot write \mathbb{R}^3 as a disjoint union of curves in such a way that each M_u is the union of a certain subset of these curves; we cannot describe the “characteristic curves of a solution u ” at all until we already know u . This might make the concept seem rather useless, but the requisite supplementary considerations will appear quite naturally when we seek an analytic description of these geometric pictures.

How would we go about finding an analytic description of the Monge cone? The Addendum to Chapter 3 suggests that the Monge cone $K(x_0, y_0, z_0)$ should be the “envelope” of the family of planes $\mathcal{F}(x_0, y_0, z_0)$; geometrically, the generators of $K(x_0, y_0, z_0)$ should be the limits of the intersections of two planes of the family $\mathcal{F}(x_0, y_0, z_0)$, the limit being formed as the two planes approach each other. Until we explicitly say the opposite, everything we now do will be based on the assumption that these limits really exist; the ensuing discussion is consequently merely a route to discovery, and does not purport to prove anything.

Let us assume for the moment that the equation

$$F(x_0, y_0, z_0, a, b) = 0$$

can be solved for b in terms of a . In other words, assume there is a function ϕ with

$$(i) \quad F(x_0, y_0, z_0, a, \phi(a)) = 0.$$

One plane of the family $\mathcal{F}(x_0, y_0, z_0)$ may be described by the equation

$$z - z_0 = a(x - x_0) + \phi(a)(y - y_0).$$

A nearby plane may be described by the equation

$$z - z_0 = (a + h)(x - x_0) + \phi(a + h)(y - y_0).$$

The points (x, y, z) in the intersection then satisfy

$$0 = h(x - x_0) + [\phi(a + h) - \phi(a)](y - y_0),$$

and hence

$$0 = (x - x_0) + \left[\frac{\phi(a + h) - \phi(a)}{h} \right] (y - y_0).$$

Therefore points in the limiting intersection ought to satisfy

$$(ii) \quad \begin{cases} z - z_0 = a(x - x_0) + \phi(a)(y - y_0) \\ 0 = (x - x_0) + \phi'(a)(y - y_0). \end{cases}$$

On the other hand, equation (i) shows that

$$\begin{aligned} 0 &= \frac{d}{da} F(x_0, y_0, z_0, a, \phi(a)) \\ &= F_p(x_0, y_0, z_0, a, \phi(a)) + \phi'(a) \cdot F_q(x_0, y_0, z_0, a, \phi(a)), \end{aligned}$$

and hence

$$(iii) \quad \phi'(a) = - \frac{F_p(x_0, y_0, z_0, a, \phi(a))}{F_q(x_0, y_0, z_0, a, \phi(a))}.$$

From (ii) and (iii) we find that the points (x, y, z) on the Monge cone $K(x_0, y_0, z_0)$ should satisfy

$$(iv) \quad \begin{cases} z - z_0 = a(x - x_0) + b(y - y_0), & \text{where } a \text{ and } b \text{ are} \\ & \text{numbers such that:} \\ F(x_0, y_0, z_0, a, b) = 0 \\ \frac{x - x_0}{F_p} = \frac{y - y_0}{F_q} & [F_p \text{ and } F_q \text{ evaluated at } (x_0, y_0, z_0, a, b)]. \end{cases}$$

Now consider a solution u of (I), and let

$$z_0 = u(x_0, y_0), \quad p_0 = u_x(x_0, y_0), \quad q_0 = u_y(x_0, y_0).$$

The tangent plane of M_u at (x_0, y_0, z_0) consists of points (x, y, z) satisfying

$$z - z_0 = p_0(x - x_0) + q_0(y - y_0).$$

Equations (iv) show that points (x, y, z) which are on both this tangent plane and the Monge cone $K(x_0, y_0, z_0)$ ought to satisfy

$$(v) \quad \frac{x - x_0}{F_p} = \frac{y - y_0}{F_q} = \frac{z - z_0}{p_0 F_p + q_0 F_q}$$

$[F_p \text{ and } F_q \text{ evaluated at } (x_0, y_0, z_0, p_0, q_0)].$

Therefore, these points ought to lie along the line through (x_0, y_0, z_0) with direction

$$(F_p, F_q, p_0 F_p + q_0 F_q) \quad [F_p \text{ and } F_q \text{ evaluated at } (x_0, y_0, z_0, p_0, q_0)].$$

We have finally reached the stage where we can make a perfectly sensible definition, involving no assumptions at all. Let u be a solution of (1), and for a point (x_0, y_0) , define z_0 , p_0 , and q_0 as before. We then define the **characteristic vector of u at (x_0, y_0)** to be the vector

$$(2) \quad X(u; x_0, y_0) = (F_p, F_q, p_0 F_p + q_0 F_q),$$

where F_p and F_q are to be evaluated at $(x_0, y_0, z_0, p_0, q_0)$; this vector is to be considered as an element of $\mathbb{R}^3_{(x_0, y_0, z_0)}$. If $M_u = \{(x, y, u(x, y))\}$, then the tangent plane of M_u at (x_0, y_0, z_0) is perpendicular to the vector $(p_0, q_0, -1)$. The vector $X(u; x_0, y_0)$ clearly has this property, so every characteristic vector of u is tangent to M_u , and the set of all characteristic vectors of u forms a vector field on M_u . The integral curves of this vector field are called the **characteristic curves of the solution u** , and they are clearly curves on M_u .

A characteristic curve c of u is thus a curve in \mathbb{R}^3 satisfying the equations

$$(3) \quad \begin{cases} \frac{dc_1(t)}{dt} = F_p(\bullet) \\ \frac{dc_2(t)}{dt} = F_q(\bullet) \\ \frac{dc_3(t)}{dt} = u_x(c_1(t), c_2(t)) \cdot F_p(\bullet) + u_y(c_1(t), c_2(t)) \cdot F_q(\bullet) \\ \text{where } \bullet = (c_1(t), c_2(t), c_3(t), u_x(c_1(t), c_2(t)), u_y(c_1(t), c_2(t))). \end{cases}$$

Now if we assume that u is C^2 , then we can also obtain equations for the partials $u_x(c_1(t), c_2(t))$ and $u_y(c_1(t), c_2(t))$. For equations (3) allow us to write

$$(4) \quad \begin{cases} \frac{du_x(c_1(t), c_2(t))}{dt} = u_{xx}(c_1(t), c_2(t)) \frac{dc_1(t)}{dt} + u_{xy}(c_1(t), c_2(t)) \frac{dc_2(t)}{dt} \\ \quad \quad \quad = u_{xx}(c_1(t), c_2(t)) F_p(\bullet) + u_{xy}(c_1(t), c_2(t)) F_q(\bullet) \\ \frac{du_y(c_1(t), c_2(t))}{dt} = u_{yx}(c_1(t), c_2(t)) F_p(\bullet) + u_{yy}(c_1(t), c_2(t)) F_q(\bullet). \end{cases}$$

On the other hand, since u satisfies

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0,$$

we also have

$$(5) \quad \begin{cases} F_x + u_x F_u + u_{xx} F_p + u_{yx} F_q = 0 \\ F_y + u_y F_u + u_{xy} F_p + u_{yy} F_q = 0, \end{cases}$$

where all partials of F are evaluated at $(x, y, u(x, y), u_x(x, y), u_y(x, y))$. Thus equations (4) become

$$(6) \quad \begin{cases} \frac{du_x(c_1(t), c_2(t))}{dt} = -F_x(\bullet) - u_x(c_1(t), c_2(t)) \cdot F_u(\bullet) \\ \frac{du_y(c_1(t), c_2(t))}{dt} = -F_y(\bullet) - u_y(c_1(t), c_2(t)) \cdot F_u(\bullet). \end{cases}$$

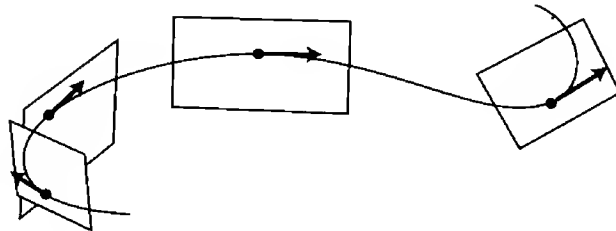
Let us now define a curve Γ in \mathbb{R}^5 by

$$(7) \quad \Gamma(t) = (c_1(t), c_2(t), c_3(t), u_x(c_1(t), c_2(t)), u_y(c_1(t), c_2(t))).$$

Then equations (3) and (6) may be written

$$(8) \quad \begin{cases} \frac{d\Gamma_1(t)}{dt} = F_p(\Gamma(t)) \\ \frac{d\Gamma_2(t)}{dt} = F_q(\Gamma(t)) \\ \frac{d\Gamma_3(t)}{dt} = \Gamma_4(t) \cdot F_p(\Gamma(t)) + \Gamma_5(t) \cdot F_q(\Gamma(t)) \\ \frac{d\Gamma_4(t)}{dt} = -F_x(\Gamma(t)) - \Gamma_4(t) \cdot F_u(\Gamma(t)) \\ \frac{d\Gamma_5(t)}{dt} = -F_y(\Gamma(t)) - \Gamma_5(t) \cdot F_u(\Gamma(t)). \end{cases}$$

Now although the curve Γ was defined in terms of a solution u , the final equations (8) involve *only* the original equation (1). This will allow us to define geometrically meaningful objects which do not depend on knowing a solution u . We may regard a point $(x_0, y_0, z_0, a, b) \in \mathbb{R}^5$ as a plane in the tangent space $\mathbb{R}^3_{(x_0, y_0, z_0)}$, namely, as the plane perpendicular to the vector $(a, b, -1)$. A curve Γ in \mathbb{R}^5 may then be regarded as a family of planes, the plane at time t being in the tangent space of \mathbb{R}^3 at $c(t) = (\Gamma_1(t), \Gamma_2(t), \Gamma_3(t))$; it will be convenient to refer to this curve c as the **base curve** of Γ . An arbitrary curve Γ is called



a **strip** if the tangent vector $c'(t)$ of the base curve c always lies in the plane determined by Γ at time t . This means that

$$c'(t) = (\Gamma_1'(t), \Gamma_2'(t), \Gamma_3'(t)) \text{ is perpendicular to } (\Gamma_4(t), \Gamma_5(t), -1).$$

So Γ is a strip if and only if it satisfies the **strip condition**:

$$(9) \quad \frac{d\Gamma_3(t)}{dt} = \Gamma_4(t) \frac{d\Gamma_1(t)}{dt} + \Gamma_5(t) \frac{d\Gamma_2(t)}{dt}.$$

Notice that any solution of (8) is automatically a strip. A curve Γ will be called a **characteristic strip** of the PDE (1) if Γ satisfies (8) and also

$$(10) \quad F(\Gamma(t)) = F(\Gamma_1(t), \Gamma_2(t), \Gamma_3(t), \Gamma_4(t), \Gamma_5(t)) = 0.$$

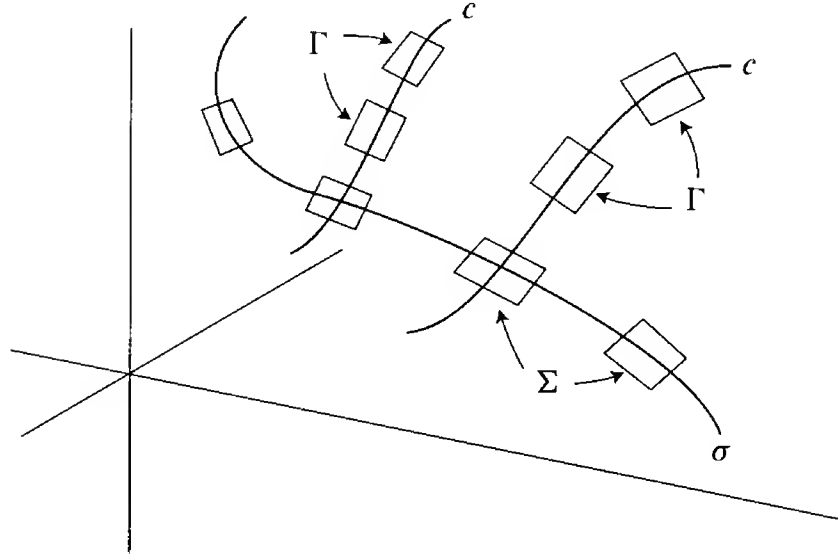
This last restriction is not as stringent as it might first seem, for if Γ satisfies (8), then

$$\begin{aligned} (11) \quad \frac{d}{dt} F(\Gamma(t)) &= F_x \frac{d\Gamma_1(t)}{dt} + \cdots + F_q \frac{d\Gamma_5(t)}{dt} \\ &\quad [\text{all partials of } F \text{ evaluated at } \Gamma(t)] \\ &= F_x F_p + F_y F_q + F_z \cdot (\Gamma_4(t) F_p + \Gamma_5(t) F_q) \\ &\quad + F_p \cdot (-F_x - \Gamma_4(t) F_z) + F_q \cdot (-F_y - \Gamma_5(t) F_z) \\ &= 0. \end{aligned}$$

So if Γ satisfies (8) and also satisfies (10) for one t , then it satisfies (10) for all t , and is consequently a characteristic strip.

Now how are characteristic strips related to solutions? We have seen that if u is a solution of (1), then M_u is the union of certain characteristic curves [solutions of (3)]. Moreover, if c is a characteristic curve, then the set of tangent planes of M_u along c gives the curve Γ of equation (7), which is a characteristic strip. So M_u is the union of base curves of characteristic strips.

Now suppose that we have an arbitrary curve Σ in \mathbb{R}^5 , with base curve σ , and that $F(\Sigma(s)) = 0$ for all s . There is a unique solution of (8) through each point $\Sigma(s)$, and by the remark after equation (11), this solution is a characteristic strip. We thus obtain a family of characteristic strips Γ . The union of the corresponding base curves c is a surface M_u , containing the base curve σ . Is it reasonable to suppose now that u is a solution of (1)? The answer is no, for there is clearly no hope unless Σ is also a strip. When this condition is satisfied, then everything works out. We will prove that if $\sigma: [a, b] \rightarrow \mathbb{R}^2$ is a given



curve, $\hat{u}: [a, b] \rightarrow \mathbb{R}$ is a given function, and $\hat{p}, \hat{q}: [a, b] \rightarrow \mathbb{R}$ are two functions satisfying

$$(a) \quad F(\Sigma(s)) = F(\sigma_1(s), \sigma_2(s), \hat{u}(s), \hat{p}(s), \hat{q}(s)) = 0,$$

and the strip condition

$$(b) \quad \frac{d\hat{u}(s)}{ds} = \hat{p}(s) \frac{d\sigma_1(s)}{ds} + \hat{q}(s) \frac{d\sigma_2(s)}{ds},$$

then there is a unique solution u of (1) satisfying

$$u(\sigma(s)) = \hat{u}(s), \quad u_x(\sigma(s)) = \hat{p}(s), \quad u_y(\sigma(s)) = \hat{q}(s)$$

[naturally, (b) is a necessary consequence of these equations]. We will clearly have to assume that $\sigma'(s)$ is linearly independent of the vector obtained by projecting the characteristic vector (2) on the (x, y) -plane. In other words, we will have to require that $\sigma'(s)$ and $(F_p(\Sigma(s)), F_q(\Sigma(s)))$ are linearly independent, or that

$$(c) \quad \sigma_1'(s) \cdot F_q(\Sigma(s)) \neq \sigma_2'(s) \cdot F_p(\Sigma(s)).$$

Before we proceed to prove the theorem, we should insert a remark about the hypotheses, which will involve σ , \hat{u} , \hat{p} , and \hat{q} satisfying (a)–(c). At first sight, we seem to be contradicting our basic philosophy about first order equations, for we seem to be saying that we can arbitrarily specify not only the values \hat{u} of u

along σ , but *also* the values $\overset{\circ}{p}$ and $\overset{\circ}{q}$ of u_x and u_y along σ . This is not really the case, for $\overset{\circ}{p}$ and $\overset{\circ}{q}$ are practically determined by the equations (a) and (b) which they must satisfy. This is most apparent when our initial curve σ is the x -axis, $\sigma(s) = (s, 0)$. Then equation (b) already determines $\overset{\circ}{p}$. Moreover, condition (c) says that $F_q \neq 0$ along $\{(s, 0, \overset{\circ}{u}(s), \overset{\circ}{p}(s), \overset{\circ}{q}(s))\}$, so the implicit function theorem shows that equation (a) can be solved for $\overset{\circ}{q}(s)$ in terms of $\overset{\circ}{p}(s)$ —there is a function ϕ with

$$F(s, 0, \overset{\circ}{u}(s), \overset{\circ}{p}(s), \phi(\overset{\circ}{p}(s))) = 0.$$

Of course, there may be several possible ϕ , but once $\overset{\circ}{q}(0)$ is determined, there will be only one continuous choice of $\overset{\circ}{q}$ satisfying (a). [In the quasi-linear case, $\overset{\circ}{q}(s)$ will actually be uniquely determined.] It is not hard to see that a similar situation prevails when σ is any curve satisfying (c): we are essentially specifying only the values $\overset{\circ}{u}$ of u along σ , and then making certain that we have a continuous choice of the limited possibilities for $\overset{\circ}{p}$ and $\overset{\circ}{q}$. In order to emphasize this point we will refer to $(\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q})$ as “initial data”, rather than as initial conditions.

3. THEOREM. Let F be a function of class C^k , $k \geq 3$, defined in an open set $U \subset \mathbb{R}^5$. Let $\sigma: [a, b] \rightarrow \mathbb{R}^2$ be a one-one C^{k-1} function, and let $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}: [a, b] \rightarrow \mathbb{R}$ be C^{k-1} functions such that for all $s \in [a, b]$ we have

$$(a) \quad \Sigma(s) = (\sigma_1(s), \sigma_2(s), \overset{\circ}{u}(s), \overset{\circ}{p}(s), \overset{\circ}{q}(s)) \in U \quad \text{and} \quad F(\Sigma(s)) = 0,$$

$$(b) \quad \frac{d\overset{\circ}{u}(s)}{ds} = \overset{\circ}{p}(s) \frac{d\sigma_1(s)}{ds} + \overset{\circ}{q}(s) \frac{d\sigma_2(s)}{ds},$$

$$(c) \quad \sigma_1'(s) \cdot F_q(\Sigma(s)) \neq \sigma_2'(s) \cdot F_p(\Sigma(s)).$$

Then there is a C^{k-1} function u , defined in a neighborhood V of $\sigma([a, b])$, which satisfies the equation

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0 \quad \text{on } V$$

and also

$$u(\sigma(s)) = \overset{\circ}{u}(s), \quad u_x(\sigma(s)) = \overset{\circ}{p}(s), \quad u_y(\sigma(s)) = \overset{\circ}{q}(s), \quad \text{for } s \in [a, b].$$

Moreover, any two functions u with this property agree on a neighborhood of $\sigma([a, b])$.

PROOF. As in the proof of Theorems 1 and 2, we use Problem I.5-5 to conclude that there is a C^{k-1} function $\alpha = (\alpha_1, \dots, \alpha_5)$ with

$$(*) \quad \begin{cases} \alpha(0, s, r) = r & \text{for } r \in \mathbb{R}^5 \\ \frac{\partial}{\partial t} \alpha_1(t, s, r) = F_p(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_2(t, s, r) = F_q(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_3(t, s, r) = \alpha_4(t, s, r) \cdot F_p(\alpha(t, s, r)) + \alpha_5(t, s, r) \cdot F_q(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_4(t, s, r) = -F_x(\alpha(t, s, r)) - \alpha_4(t, s, r) \cdot F_u(\alpha(t, s, r)) \\ \frac{\partial}{\partial t} \alpha_5(t, s, r) = -F_y(\alpha(t, s, r)) - \alpha_5(t, s, r) \cdot F_u(\alpha(t, s, r)). \end{cases}$$

Let

$$\beta(s, t) = \alpha(t, s, \sigma_1(s), \sigma_2(s), \dot{u}(s), \dot{p}(s), \dot{q}(s)),$$

so that β is also C^{k-1} . In particular,

$$\beta(s, 0) = (\sigma_1(s), \sigma_2(s), \dot{u}(s), \dot{p}(s), \dot{q}(s)) = \Sigma(s).$$

If we define

$$\gamma(s, t) = (\beta_1(s, t), \beta_2(s, t)) \in \mathbb{R}^2,$$

then the Jacobian of γ at $(s, 0)$ is

$$\begin{aligned} \begin{pmatrix} \frac{\partial \beta_1}{\partial s}(s, 0) & \frac{\partial \beta_1}{\partial t}(s, 0) \\ \frac{\partial \beta_2}{\partial s}(s, 0) & \frac{\partial \beta_2}{\partial t}(s, 0) \end{pmatrix} &= \begin{pmatrix} \sigma_1'(s) & \frac{\partial \alpha_1}{\partial t}(0, s, \Sigma(s)) \\ \sigma_2'(s) & \frac{\partial \alpha_2}{\partial t}(0, s, \Sigma(s)) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1'(s) & F_p(\Sigma(s)) \\ \sigma_2'(s) & F_q(\Sigma(s)) \end{pmatrix} \quad \text{by } (*), \end{aligned}$$

and this is non-singular by hypothesis. So if ε is sufficiently small, then $\gamma: [a, b] \times (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^2$ is a C^{k-1} diffeomorphism onto a neighborhood V of $\sigma([a, b])$.

The solution u , if it exists, clearly must be the C^{k-1} function

$$u(x, y) = \beta_3(\gamma^{-1}(x, y)) \quad \text{or equivalently} \quad u(\gamma(s, t)) = \beta_3(s, t).$$

We claim that

$$u_x(\gamma(s, t)) = \beta_4(s, t) \quad \text{and} \quad u_y(\gamma(s, t)) = \beta_5(s, t).$$

This will prove that

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0;$$

for we have already seen (equation 11) that $F(\alpha(t, s, r))$ is constant for fixed s and r , while $F(\alpha(0, s, \Sigma(s))) = 0$ by (a), so that we will have

$$\begin{aligned} 0 &= F(\alpha(t, s, \Sigma(s))) = F(\beta(s, t)) = F(\beta_1(s, t), \beta_2(s, t), \beta_3(s, t), \beta_4(s, t), \beta_5(s, t)) \\ &= F(\gamma(s, t), u(\gamma(s, t)), u_x(\gamma(s, t)), u_y(\gamma(s, t))). \end{aligned}$$

To prove the claim, we consider the function

$$\Delta = \frac{\partial \beta_3}{\partial s} - \beta_4 \cdot \frac{\partial \beta_1}{\partial s} - \beta_5 \cdot \frac{\partial \beta_2}{\partial s}.$$

We have

$$\begin{aligned} \Delta(s, 0) &= \frac{d\mathring{u}(s)}{ds} - \mathring{p}(s) \cdot \frac{d\sigma_1(s)}{ds} - \mathring{q}(s) \cdot \frac{d\sigma_2(s)}{ds} \\ &= 0 \quad \text{by (b).} \end{aligned}$$

Moreover,

$$\begin{aligned} \frac{\partial \Delta}{\partial t} &= \frac{\partial^2 \beta_3}{\partial s \partial t} - \frac{\partial \beta_4}{\partial t} \frac{\partial \beta_1}{\partial s} - \frac{\partial \beta_5}{\partial t} \frac{\partial \beta_2}{\partial s} - \beta_4 \cdot \frac{\partial^2 \beta_1}{\partial s \partial t} - \beta_5 \cdot \frac{\partial^2 \beta_2}{\partial s \partial t} \\ &= \frac{\partial}{\partial s} \left(\frac{\partial \beta_3}{\partial t} - \beta_4 \cdot \frac{\partial \beta_1}{\partial t} - \beta_5 \cdot \frac{\partial \beta_2}{\partial t} \right) \\ &\quad + \frac{\partial \beta_4}{\partial s} \frac{\partial \beta_1}{\partial t} + \frac{\partial \beta_5}{\partial s} \frac{\partial \beta_2}{\partial t} - \frac{\partial \beta_4}{\partial t} \frac{\partial \beta_1}{\partial s} - \frac{\partial \beta_5}{\partial t} \frac{\partial \beta_2}{\partial s} \\ &= 0 + F_p \cdot \frac{\partial \beta_4}{\partial s} + F_q \cdot \frac{\partial \beta_5}{\partial s} + (F_x + F_u \beta_4) \frac{\partial \beta_1}{\partial s} + (F_y + F_u \beta_5) \frac{\partial \beta_2}{\partial s} \\ &\quad \text{by (*)} \quad [\text{where all partials of } F \text{ are evaluated at } \beta(s, t)] \\ &= F_x \cdot \frac{\partial \beta_1}{\partial s} + F_y \cdot \frac{\partial \beta_2}{\partial s} + F_u \cdot \frac{\partial \beta_3}{\partial s} + F_p \cdot \frac{\partial \beta_4}{\partial s} + F_q \cdot \frac{\partial \beta_5}{\partial s} \\ &\quad - F_u \cdot \left(\frac{\partial \beta_3}{\partial s} - \beta_4 \cdot \frac{\partial \beta_1}{\partial s} - \beta_5 \cdot \frac{\partial \beta_2}{\partial s} \right) \\ &= \frac{\partial}{\partial s} (F(\beta(s, t))) - F_u \cdot \Delta \\ &= -F_u \cdot \Delta, \end{aligned}$$

since we have already seen that $F(\beta(s, t)) = 0$. Now for each fixed s , we have an ordinary differential equation

$$\frac{\partial \Delta}{\partial t} = -F_u \cdot \Delta,$$

with the initial condition

$$\Delta(s, 0) = 0,$$

so the unique solution is $\Delta(s, t) = 0$. In other words, we have shown that

$$\frac{\partial \beta_3}{\partial s} = \beta_4 \cdot \frac{\partial \beta_1}{\partial s} + \beta_5 \cdot \frac{\partial \beta_2}{\partial s}.$$

Also

$$\frac{\partial \beta_3}{\partial t} = \beta_4 \cdot \frac{\partial \beta_1}{\partial t} + \beta_5 \cdot \frac{\partial \beta_2}{\partial t} \quad \text{by } (*).$$

On the other hand, differentiating the definition $u(\gamma(s, t)) = \beta_3(s, t)$ gives

$$\begin{aligned} \frac{\partial \beta_3}{\partial s} &= u_x(\gamma(s, t)) \cdot \frac{\partial \beta_1}{\partial s} + u_y(\gamma(s, t)) \cdot \frac{\partial \beta_2}{\partial s} \\ \frac{\partial \beta_3}{\partial t} &= u_x(\gamma(s, t)) \cdot \frac{\partial \beta_1}{\partial t} + u_y(\gamma(s, t)) \cdot \frac{\partial \beta_2}{\partial t}. \end{aligned}$$

These last four equations give two solutions for two linear equations in two unknowns, whose determinant

$$\det \begin{pmatrix} \frac{\partial \beta_1}{\partial s} & \frac{\partial \beta_2}{\partial s} \\ \frac{\partial \beta_1}{\partial t} & \frac{\partial \beta_2}{\partial t} \end{pmatrix}$$

is $\neq 0$ for $(s, t) \in [a, b] \times (-\varepsilon, \varepsilon)$. So the two solutions must be the same, i.e.,

$$u_x(\gamma(s, t)) = \beta_4(s, t) \quad \text{and} \quad u_y(\gamma(s, t)) = \beta_5(s, t),$$

as desired. ♦

We will say that the initial curve σ is **free for the initial data** $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}$ when condition (c) in Theorem 3 is satisfied. Thus we can solve the Cauchy problem for a first order PDE (I) for any initial strip $\Sigma = (\sigma_1, \sigma_2, \overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q})$ for which the initial curve σ is free for the initial data $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}$.

Again we consider the case where our initial curve σ fails to be free for the initial data $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}$ in the worst possible way, namely when $\sigma'(s)$ and $(F_p(\Sigma(s)), F_q(\Sigma(s)))$ are everywhere linearly dependent. Once again we say that σ is **characteristic** for $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}$. Assuming that σ is an imbedding, we can reparameterize σ so that $\sigma'(s) = (F_p(\Sigma(s)), F_q(\Sigma(s)))$. This gives us the first two equations in (8) for the curve $(\sigma_1, \sigma_2, \overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q})$. The third equation of (8) is just the strip condition (b). The argument on pages 18–19 shows that these three equations imply the last two if there is a solution u of (I) with

$$u(\sigma(s)) = \overset{\circ}{u}(s), \quad u_x(\sigma(s)) = \overset{\circ}{p}(s), \quad u_y(\sigma(s)) = \overset{\circ}{q}(s).$$

So when σ is characteristic, the Cauchy problem is solvable for the initial data $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}$ along σ only if $(\sigma_1, \sigma_2, \overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q})$ is a characteristic strip. When this is the case, there will be infinitely many solutions with this initial data along σ . The base curves of characteristic strips are the intersection curves of the graphs of different solutions meeting tangentially.

We can now describe the situation for first order PDE's in n variables very easily, without bothering to write down all the results as formal theorems. Consider first the quasi-linear PDE

$$\sum_{i=1}^n A_i(x_1, \dots, x_n, u) \cdot u_{x_i} = C(x_1, \dots, x_n, u).$$

The **characteristic vector field** of this equation is the vector field X in \mathbb{R}^{n+1} defined by

$$X = \sum_{i=1}^n A_i \frac{\partial}{\partial x_i} + C \frac{\partial}{\partial z};$$

the integral curves of X are the **characteristic curves** of the equation. As in the case $n = 2$, it is clear that if u is a solution of (I), then the hypersurface

$$M_u = \{(x_1, \dots, x_n, u(x_1, \dots, x_n))\} \subset \mathbb{R}^{n+1}$$

is a union of characteristic curves. Now suppose we are given a one-one map

$$\sigma: \mathcal{D} \rightarrow \mathbb{R}^n,$$

where $\mathcal{D} \subset \mathbb{R}^{n-1}$ is a compact $(n-1)$ -dimensional manifold-with-boundary, and a function $\overset{\circ}{u}: \mathcal{D} \rightarrow \mathbb{R}$. We can produce a solution u of (I) with

$$u(\sigma(s)) = \overset{\circ}{u}(s) \quad \text{for all } s \in \mathcal{D}$$

by taking the union of the characteristic curves through all points $(\sigma(s), \mathring{u}(s)) \in \mathbb{R}^{n+1}$. The proof is exactly analogous to the proof of Theorem 2, except that we will now require that the matrix

$$\begin{pmatrix} D_1\sigma_1(s) & \dots & D_{n-1}\sigma_1(s) & A_1(\sigma(s), \mathring{u}(s)) \\ \vdots & & \vdots & \vdots \\ D_1\sigma_n(s) & \dots & D_{n-1}\sigma_n(s) & A_n(\sigma(s), \mathring{u}(s)) \end{pmatrix}$$

be non-singular for all $s \in \mathcal{D}$. This means, first of all, that the matrix $(D_j\sigma_i(s))$ must have rank $n - 1$, so that σ is an imbedding and $\sigma(\mathcal{D}) \subset \mathbb{R}^n$ is a hypersurface. In addition, the vector $(A_1(\sigma(s), \mathring{u}(s)), \dots, A_n(\sigma(s), \mathring{u}(s)))$ must not lie in the tangent space of $\sigma(\mathcal{D})$; we express this by saying that the “initial manifold” $\sigma(\mathcal{D})$ is **free for the initial condition** \mathring{u} (for linear equations the initial condition \mathring{u} is irrelevant). Thus we can solve the Cauchy problem for any initial condition along an initial $(n - 1)$ -manifold which is free for the initial condition.

Now we consider the general first order PDE

$$F(x_1, \dots, x_n, u(x_1, \dots, x_n), u_{x_1}(x_1, \dots, x_n), \dots, u_{x_n}(x_1, \dots, x_n)) = 0.$$

We denote the partials of F by

$$F_{x_i}, \quad F_u, \quad F_{p_i}.$$

Consider curves Γ in \mathbb{R}^{2n+1} satisfying

$$\begin{cases} \frac{d\Gamma_i(t)}{dt} = F_{p_i}(\Gamma(t)) & i = 1, \dots, n \\ \frac{d\Gamma_{n+1}(t)}{dt} = \sum_{i=1}^n \Gamma_{n+1+i}(t) \cdot F_{p_i}(\Gamma(t)) \\ \frac{d\Gamma_{n+1+i}(t)}{dt} = -F_{x_i}(\Gamma(t)) - \Gamma_{n+1+i}(t) F_u(\Gamma(t)) & i = 1, \dots, n. \end{cases}$$

As before, we easily check that if Γ satisfies these equations, then $F(\Gamma(t))$ is constant in t . A solution Γ with $F(\Gamma(t)) = 0$ for all t is called a **characteristic strip**. Now suppose we have a one-one map

$$\sigma: \mathcal{D} \rightarrow \mathbb{R}^n$$

with $\mathcal{D} \subset \mathbb{R}^{n-1}$ as before, and functions

$$\mathring{u}, \mathring{p}_1, \dots, \mathring{p}_n: \mathcal{D} \rightarrow \mathbb{R}$$

with

$$F(\Sigma(s)) = F(\sigma_1(s), \dots, \sigma_n(s), \overset{\circ}{u}(s), \overset{\circ}{p}_1(s), \dots, \overset{\circ}{p}_n(s)) = 0 \quad \text{for all } s \in \mathcal{D}.$$

Then there is a unique characteristic strip Γ through each point $\Sigma(s)$, and the union of the corresponding base curves is a hypersurface M_u . In order for the function u to be a solution to our PDE we will need two conditions, which allow us to extend the proof of Theorem 3 essentially without change. First, the matrix

$$\begin{pmatrix} D_1\sigma_1(s) & \dots & D_{n-1}\sigma_1(s) & F_{p_1}(\Sigma(s)) \\ \vdots & & \vdots & \vdots \\ D_1\sigma_n(s) & \dots & D_{n-1}\sigma_n(s) & F_{p_n}(\Sigma(s)) \end{pmatrix}$$

must be non-singular. Thus $\sigma(\mathcal{D}) \subset \mathbb{R}^n$ must be an $(n-1)$ -manifold, and $(F_{p_1}(\Sigma(s)), \dots, F_{p_n}(\Sigma(s)))$ must not lie in its tangent space—once again, we express this by saying that the initial manifold $\sigma(\mathcal{D})$ is **free for the initial data** $\overset{\circ}{u}, \overset{\circ}{p}_1, \dots, \overset{\circ}{p}_n$. Second, we must have

$$\frac{\partial \overset{\circ}{u}}{\partial s_j} = \sum_{i=1}^n \overset{\circ}{p}_i \cdot \frac{\partial \sigma_i}{\partial s_j}.$$

In terms of Σ , this condition reads

$$\frac{\partial \Sigma_{n+1}}{\partial s_j} = \sum_{i=1}^n \Sigma_{n+1+i} \cdot \frac{\partial \Sigma_i}{\partial s_j},$$

and is called the **strip manifold condition**. If we think of a point $(x_1, \dots, x_n, z, p_1, \dots, p_n)$ in \mathbb{R}^{2n+1} as a hyperplane in $\mathbb{R}^{n+1}_{(x_1, \dots, x_n, z)}$, namely as the hyperplane perpendicular to the vector $(p_1, \dots, p_n, -1)$, then $\Sigma: \mathcal{D} \rightarrow \mathbb{R}^{2n+1}$ may be regarded as a family of hyperplanes along the $(n-1)$ -dimensional submanifold $\sigma(\mathcal{D})$. It is easy to see that Σ satisfies the strip manifold condition if and only if the tangent space of $\sigma(\mathcal{D})$ at any point $\sigma(s)$ always lies in the hyperplane determined by Σ at s . We may summarize by saying that we can solve the Cauchy problem for any strip manifold $(\sigma_1, \dots, \sigma_n, \overset{\circ}{u}, \overset{\circ}{p}_1, \dots, \overset{\circ}{p}_n)$ for which the initial $(n-1)$ -dimensional submanifold $\sigma(\mathcal{D})$ is free for the initial data $\overset{\circ}{u}, \overset{\circ}{p}_1, \dots, \overset{\circ}{p}_n$.

2. FREE INITIAL MANIFOLDS FOR HIGHER ORDER EQUATIONS

In the previous section we found that the characteristic curves or characteristic strips for first order equations were the clue to solving them, while the free hypersurfaces were the appropriate initial manifolds for which we could solve the Cauchy problem. For higher order equations things are not nearly so simple, but we can at least decide at the outset what the free initial manifolds ought to be. To do this, we will first consider the special case where the initial manifold is $M = \{x \in \mathbb{R}^n : x_n = 0\} \subset \mathbb{R}^n$.

First a review of the situation for first order equations. For the quasi-linear equation

$$\sum_{i=1}^n A_i(x_1, \dots, x_n, u) \cdot u_{x_i} = C(x_1, \dots, x_n, u),$$

the manifold $M = \{x \in \mathbb{R}^n : x_n = 0\}$ is free for the initial condition $\overset{\circ}{u}$ on M if and only if

$$A_n(x_1, \dots, x_{n-1}, 0, \overset{\circ}{u}(x_1, \dots, x_{n-1})) \neq 0 \quad \text{on } M.$$

If this condition holds, then in a neighborhood of M we can write our equation as an equation for u_{x_n} in terms of u and the other u_{x_i} :

$$\begin{aligned} u_{x_n} &= - \sum_{i=1}^{n-1} \frac{A_i(x_1, \dots, x_n, u)}{A_n(x_1, \dots, x_n, u)} u_{x_i} + \frac{C(x_1, \dots, x_n, u)}{A_n(x_1, \dots, x_n, u)} \\ &= f(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_{n-1}}), \end{aligned}$$

where the function f is defined in a neighborhood of all points

$$(x_1, \dots, x_{n-1}, 0, \overset{\circ}{u}(x_1, \dots, x_{n-1}), p_1, \dots, p_{n-1}).$$

On the other hand, if M fails to be free at some point, then our original equation gives us a relationship between the u_{x_i} for $i < n$, which means that there are additional conditions which $\overset{\circ}{u}$ would have to satisfy for a solution to exist.

For the general first order PDE

$$F(x, u(x), u_{x_1}(x), \dots, u_{x_n}(x)) = 0 \quad [x = (x_1, \dots, x_n)],$$

the initial data $\overset{\circ}{u}, \overset{\circ}{p}_1, \dots, \overset{\circ}{p}_n$ must satisfy, using $x_{1\dots n-1}$ as an abbreviation for x_1, \dots, x_{n-1} ,

$$\begin{aligned} (a) \quad 0 &= F(x_{1\dots n-1}, 0, \overset{\circ}{u}(x_{1\dots n-1}), \overset{\circ}{p}_1(x_{1\dots n-1}), \dots, \overset{\circ}{p}_n(x_{1\dots n-1})) \\ &= F(\Sigma(s)) \quad s = (x_1, \dots, x_{n-1}), \end{aligned}$$

as well as the obvious compatibility conditions

$$(b) \quad \frac{\partial \overset{\circ}{u}}{\partial x_j} = \overset{\circ}{p}_j \quad j = 1, \dots, n-1$$

which is what the strip manifold condition on page 28 boils down to in this case; otherwise expressed, the initial value $\overset{\circ}{u}$ of u along M already determines the values of u_{x_i} along M for $i < n$, so the only other initial data that we need is a value $\overset{\circ}{p}_n$ of u_{x_n} along M satisfying (a), when the $\overset{\circ}{p}_j$ for $j < n$ are defined by (b). Now M is free for this initial data if and only if

$$F_{p_n}(\Sigma(s)) \neq 0 \quad \text{on } M.$$

In this case, the implicit function theorem tells us that there is a unique function f defined in a neighborhood of any given

$$\bullet = (x_1, \dots, x_{n-1}, 0, \overset{\circ}{u}(x_1, \dots, x_{n-1}), \overset{\circ}{p}_1(x_1, \dots, x_{n-1}), \dots, \overset{\circ}{p}_{n-1}(x_1, \dots, x_{n-1}))$$

such that

$$\begin{cases} F(x, z, p_1, \dots, p_{n-1}, f(x, z, p_1, \dots, p_{n-1})) = 0 & \text{in this neighborhood} \\ f(\bullet) = \overset{\circ}{p}_n(x_1, \dots, x_{n-1}). \end{cases}$$

So our PDE is equivalent to the equation

$$u_{x_n} = f(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_{n-1}})$$

expressing u_{x_n} in terms of u and the other u_{x_i} . On the other hand, if M is not free at some point, so that $F_{p_n}(\Sigma(s_0)) = 0$ for some s_0 , then we generally cannot find any continuous initial data $\overset{\circ}{p}_n$ satisfying (a).

We will now generalize this discussion to decide when $M = \{x \in \mathbb{R}^n : x_n = 0\}$ should be called free for a second order equation. First consider the quasi-linear second order equation

$$(l) \quad \sum_{i,j=1}^n A_{ij} u_{x_i x_j} = C,$$

where the functions A_{ij} and C depend not only on x_1, \dots, x_n , and u , but also on the u_{x_i} . It seems reasonable that the initial conditions for the Cauchy problem should be the values $\overset{\circ}{u}, \overset{\circ}{p}_1, \dots, \overset{\circ}{p}_n$ of u and its first derivatives on M . But, as we have already noted, $\overset{\circ}{u}$ already determines the $\overset{\circ}{p}_i$ for $i < n$. So the initial

conditions for the Cauchy problem* should be the values $\overset{\circ}{u}, \overset{\circ}{p}_n$ of u and u_{x_n} along M . For the PDE (1) we will define M to be **free for the initial conditions** $\overset{\circ}{u}, \overset{\circ}{p}_n$ if (recall that we are using $x_{1\dots n-1}$ as an abbreviation for x_1, \dots, x_{n-1})

$$\begin{aligned} A_{nn}(\Sigma(s)) &= A_{nn}(x_{1\dots n-1}, 0, \overset{\circ}{u}(x_{1\dots n-1}), \overset{\circ}{p}_1(x_{1\dots n-1}), \dots, \overset{\circ}{p}_n(x_{1\dots n-1})) \\ &\neq 0 \quad \text{on } M, \end{aligned}$$

where $\overset{\circ}{p}_i = \partial \overset{\circ}{u} / \partial x_i$ for $i < n$. If this condition holds, then in a neighborhood of any point of M we can write our equation as an equation for $u_{x_n x_n}$ in terms of u , the first partials u_{x_i} , and the other second partials $u_{x_i x_j}$:

$$\begin{aligned} u_{x_n x_n} &= - \sum_{\substack{i,j=1 \\ (i,j) \neq (n,n)}}^n \frac{A_{ij}}{A_{nn}} u_{x_i x_j} + \frac{C}{A_{nn}} \\ &= f(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_n}, \dots, u_{x_i x_j}, \dots), \\ &\quad [(i, j) \neq (n, n)] \end{aligned}$$

where the function f is defined in a neighborhood of all points

$$(x_{1\dots n-1}, 0, \overset{\circ}{u}(x_{1\dots n-1}), \overset{\circ}{p}_1(x_{1\dots n-1}), \dots, \overset{\circ}{p}_n(x_{1\dots n-1}), \dots, p_{ij}(x_{1\dots n-1}), \dots).$$

Now consider the general second order equation

$$(2) \quad F(x, u(x), \dots, u_{x_i}(x), \dots, u_{x_i x_j}(x), \dots) = 0.$$

Appropriate initial data will be functions

$$\overset{\circ}{u}, \quad \overset{\circ}{p}_i, \quad \overset{\circ}{p}_{ij}$$

satisfying

$$\begin{aligned} (a) \quad 0 &= F(x_{1\dots n-1}, 0, \overset{\circ}{u}(x_{1\dots n-1}), \dots, \overset{\circ}{p}_i(x_{1\dots n-1}), \dots, \overset{\circ}{p}_{ij}(x_{1\dots n-1}), \dots) \\ &= F(\Sigma(s)), \end{aligned}$$

*To avoid any confusion about the basic philosophy of the Cauchy problem, we emphasize that for a *quasi-linear second order* equation, the “initial conditions” $\overset{\circ}{u}, \overset{\circ}{p}_n$ are completely arbitrary, while for the *general first order* equation the “initial data” $\overset{\circ}{p}_n$ must satisfy (a) on page 29, and are essentially uniquely determined by the value at a single point—we have to include $\overset{\circ}{p}_n$ in the initial data just to show which of the possible solutions of (a) we are considering.

and

$$(b) \quad \begin{cases} \frac{\partial \overset{\circ}{u}}{\partial x_j} = p_{j0} & j < n \\ \frac{\partial \overset{\circ}{p}_i}{\partial x_j} = \overset{\circ}{p}_{ij} & i \leq n, j < n; \end{cases}$$

in other words, we really need only $\overset{\circ}{u}, \overset{\circ}{p}_n, \overset{\circ}{p}_{nn}$ satisfying (a), when the other $\overset{\circ}{p}_j$ and $\overset{\circ}{p}_{ij}$ are defined by (b). For the PDE (2) we define M to be **free for the initial data** $\overset{\circ}{u}, \overset{\circ}{p}_n, \overset{\circ}{p}_{nn}$ if

$$F_{p_{nn}}(\Sigma(s)) \neq 0 \quad \text{on } M.$$

In this case there is a unique function f defined in a neighborhood of any given point

$$\bullet = (x_{1\dots n-1}, 0, \overset{\circ}{u}(x_{1\dots n-1}), \dots, \overset{\circ}{p}_i(x_{1\dots n-1}), \dots, \overset{\circ}{p}_{ij}(x_{1\dots n-1}), \dots) \\ [(i, j) \neq (n, n)]$$

such that

$$\begin{cases} F(x, z, \dots p_i \dots p_{ij} \dots f(x, z, \dots p_i \dots p_{ij} \dots)) = 0 & \text{in this neighborhood} \\ f(\bullet) = \overset{\circ}{p}_{nn}(x_1, \dots, x_{n-1}). \end{cases}$$

So our PDE is equivalent to the equation

$$u_{x_n x_n} = f(x_1, \dots, x_n, u, \dots, u_{x_i}, \dots, u_{x_i x_j}, \dots) \quad [(i, j) \neq (n, n)]$$

expressing $u_{x_n x_n}$ in terms of u , the partials u_{x_i} , and the other second partial derivatives $u_{x_i x_j}$.

Now we are ready to decide when an arbitrary $(n-1)$ -dimensional submanifold $M \subset \mathbb{R}^n$ should be called free for a second order PDE. Again we begin with the quasi-linear equation

$$(I) \quad \sum_{i,j=1}^n A_{ij} u_{x_i x_j} = C.$$

It seems reasonable that the initial conditions for the Cauchy problem on M should be the value $\overset{\circ}{u}$ of u on M , together with the **normal derivative** u' of u on M ,

$$u'(p) = \lim_{h \rightarrow 0} \frac{u(p + h \cdot v(p)) - u(p)}{h},$$

where $\nu: M \rightarrow \mathbb{R}^n$ is the unit normal on M . This normal derivative is the same as

$$u'(p) = \nabla_{\nu(p)} u,$$

where ∇ denotes the ordinary covariant derivative in \mathbb{R}^n . From the value $\overset{\circ}{u}$ of u on M we can calculate any directional derivative $\nabla_{X_p} u$ for which X_p is tangent to M at p . So from $\overset{\circ}{u}$ and u' we can calculate *all* directional derivatives $\nabla_{Y_p} u$, for $p \in M$ and $Y_p \in \mathbb{R}^n_p$.

Now choose a diffeomorphism $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\phi(M) \subset \{x \in \mathbb{R}^n : x^n = 0\}$. We look for a solution of (I) of the form $u = \tilde{u} \circ \phi$. Substituting the expressions

$$(*) \quad \begin{cases} u_{x_i} = \sum_k \tilde{u}_{x_k} \cdot \phi^k_{x_i} \\ u_{x_i x_j} = \sum_{k,l} \tilde{u}_{x_k x_l} \phi^k_{x_i} \phi^l_{x_j} + \sum_k \tilde{u}_{x_k} \phi^k_{x_i x_j} \end{cases}$$

into (I), we obtain a quasi-linear equation for \tilde{u} ,

$$(I') \quad \sum_{i,j=1}^n \tilde{A}_{ij} \tilde{u}_{x_i x_j} = \tilde{C}.$$

Prescribing the value $\overset{\circ}{u}$ of u on M is equivalent to prescribing the value $\overset{\circ}{\tilde{u}}$ of \tilde{u} on $\{x \in \mathbb{R}^n : x^n = 0\}$. If we also know u' on M , then we know all directional derivatives of u on M , and consequently all directional derivatives of \tilde{u} on $\{x \in \mathbb{R}^n : x^n = 0\}$; in particular, we know $\partial \tilde{u} / \partial x_n$. So solving the Cauchy problem for (I) for the initial conditions $\overset{\circ}{u}, u'$ on M is equivalent to solving the Cauchy problem for (I') for given initial conditions $\overset{\circ}{\tilde{u}}, \overset{\circ}{p}_n$ on $\{x \in \mathbb{R}^n : x^n = 0\}$. Now we ask: what conditions on $(M, \overset{\circ}{u}, u')$ will make $\{x \in \mathbb{R}^n : x^n = 0\}$ free for the initial conditions $\overset{\circ}{\tilde{u}}, \overset{\circ}{p}_n$ for equation (I')? In other words, when will the coefficient \tilde{A}_{nn} of $\tilde{u}_{x_n x_n}$ in (I') be non-zero? From the derivation of equation (I') we see immediately that

$$\tilde{A}_{nn} = \sum_{i,j} A_{ij} \phi^n_{x_i} \phi^n_{x_j}.$$

Since $M = (\phi^n)^{-1}(0)$, the vector $(\phi^n_{x_1}, \dots, \phi^n_{x_n})$ is a multiple of the normal ν of M (compare pg. II.113). So for the PDE (I) we define the $(n-1)$ -dimensional submanifold $M \subset \mathbb{R}^n$ to be **free for the initial conditions** $\overset{\circ}{u}, u'$ if

$$\sum_{i,j} A_{ij} \nu_i \nu_j \neq 0 \quad \text{on } M$$

(in order to write this equation out for a point of M , we have to know the

values of u and the u_{x_i} at this point, since these occur as arguments of A_{ij} ; but we can compute these from the initial values $\overset{\circ}{u}, u'$. If M is free for the initial conditions $\overset{\circ}{u}, u'$, then equation (1) with the initial conditions $\overset{\circ}{u}, u'$ is equivalent to an equation for a function \tilde{u} expressing $\tilde{u}_{x_n x_n}$ in terms of \tilde{u} , first partials of \tilde{u} , and the remaining second partials of \tilde{u} , with initial conditions giving the values of \tilde{u} and \tilde{u}_{x_n} at points $(x_1, \dots, x_{n-1}, 0)$.

Now consider the general second order PDE

$$(2) \quad F(x, u(x), \dots, u_{x_i}(x), \dots, u_{x_i x_j}(x), \dots) = 0.$$

Appropriate initial data for the Cauchy problem on an $(n - 1)$ -dimensional submanifold $M \subset \mathbb{R}^n$ will be functions

$$\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij},$$

giving the values of u and its first and second partial derivatives on M . Of course, the $\overset{\circ}{p}_i$ can be determined by giving the normal derivative u' along M , which can be prescribed arbitrarily. But the $\overset{\circ}{p}_{ij}$ must satisfy

$$(a) \quad 0 = F(x, \overset{\circ}{u}(x), \dots, \overset{\circ}{p}_i(x), \dots, \overset{\circ}{p}_{ij}(x), \dots) = 0 \quad \text{for } x \in M,$$

as well as certain compatibility conditions; if M is the image of the map

$$(s_1, \dots, s_{n-1}) \mapsto (\sigma_1(s_1, \dots, s_{n-1}), \dots, \sigma_n(s_1, \dots, s_{n-1})) \in \mathbb{R}^n,$$

and we regard $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$ as functions of (s_1, \dots, s_{n-1}) , then these conditions can be written as

$$(b) \quad \begin{cases} \frac{\partial \overset{\circ}{u}}{\partial s_j} = \sum_{i=1}^n \overset{\circ}{p}_i \cdot \frac{\partial \sigma_i}{\partial s_j} \\ \frac{\partial \overset{\circ}{p}_i}{\partial s_j} = \sum_{i=1}^n \overset{\circ}{p}_{ij} \cdot \frac{\partial \sigma_i}{\partial s_j}. \end{cases}$$

Once again, choose a diffeomorphism $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\phi(M) \subset \{x \in \mathbb{R}^n : x^n = 0\}$, and consider a solution of (2) of the form $u = \tilde{u} \circ \phi$. Equation (2) is equivalent to a second order PDE for \tilde{u}

$$(2') \quad \tilde{F}(x, \tilde{u}(x), \dots, \tilde{u}_{x_i}(x), \dots, \tilde{u}_{x_i x_j}(x), \dots) = 0,$$

and prescribing the functions $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$ on M is equivalent to prescribing functions $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$ on $\{x \in \mathbb{R}^n : x_n = 0\}$ which satisfy the conditions (a) and (b) for the equation (2') on this initial manifold. We want to know when $\{x \in \mathbb{R}^n : x_n = 0\}$ will be free for this initial data; thus we want to know when $\tilde{F}_{\tilde{p}_{nn}}$, evaluated at suitable points, is non-zero. Since we get equation (2') by substituting (*) into (2), we easily see that

$$\tilde{F}_{\tilde{p}_{nn}} = \sum_{i,j=1}^n F_{p_{ij}} \phi^n_{x_i} \phi^n_{x_j}.$$

So for the PDE (2) we define M to be **free for the initial data** $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$ if

$$\sum_{i,j} F_{p_{ij}} v_i v_j \neq 0 \quad \text{on } M$$

(in order to write this equation out for a point of M we need to know the values of $u, u_{x_i}, u_{x_i x_j}$ on M , which are given to us). If M is free for the data $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$, then the Cauchy problem for the general second order PDE (2) is equivalent to the Cauchy problem for an equation for a function \tilde{u} expressing $\tilde{u}_{x_n x_n}$ in terms of \tilde{u} , first partials of \tilde{u} , and the remaining second partials of \tilde{u} .

As in the case of first order equations we define M to be **characteristic** for the initial data $\overset{\circ}{u}, \overset{\circ}{p}_i, \overset{\circ}{p}_{ij}$ if M fails to be free in the worst possible way, that is, if

$$\sum_{i,j} F_{p_{ij}} v_i v_j = 0 \quad \text{on } M.$$

Since we will never consider PDE's of order higher than 2, we will not bother to carry out a similar discussion for these equations. We merely note that with the appropriate definitions, solving the Cauchy problem for a k^{th} order PDE when the initial manifold is free for the initial data is always equivalent to solving an equation

$$\frac{\partial^k u}{\partial x_n^k}(x) = f \left(x, u(x), \dots, \frac{\partial^l u}{\partial x_1^{i_1} \dots \partial x_n^{i_n}}, \dots \right),$$

in which the order i_n of any partial on the right with respect to x_n is $\leq k - 1$, with initial conditions for

$$u(x_1, \dots, x_{n-1}, 0), \dots, \frac{\partial^{k-1} u}{\partial x_n^{k-1}}(x_1, \dots, x_{n-1}, 0).$$

3. SYSTEMS OF FIRST ORDER PDE's

For an ordinary differential equation

$$(1) \quad u'(x) = f(x, u(x)),$$

we found that the existence of solutions was no harder to prove for the case of a function $u: \mathbb{R} \rightarrow \mathbb{R}^n$ than it was for the case of a function $u: \mathbb{R} \rightarrow \mathbb{R}$. So we could consider (1) to be a *system* of equations

$$u_i'(x) = f_i(x, u_1(x), \dots, u_n(x)).$$

This enabled us to solve an n^{th} order equation

$$(2) \quad u^{(n)}(x) = f(x, u(x), u'(x), \dots, u^{(n-1)}(x)),$$

for equation (2) is equivalent to the system of equations

$$(3) \quad \begin{cases} u' = u_1 \\ u_1' = u_2 \\ \vdots \\ u_{n-2}' = u_{n-1} \\ u_{n-1}'(x) = f(x, u(x), \dots, u_{n-1}(x)). \end{cases}$$

More precisely, if u satisfies (2), then $(u, u', \dots, u^{(n-1)})$ satisfies (3); conversely, if (u, u_1, \dots, u_{n-1}) satisfies (3), then u satisfies (2) [and moreover $u_i = u^{(i)}$]. Since (3) can be solved with any initial conditions $(u(x_0), \dots, u_{n-1}(x_0))$, equation (2) can be solved with any initial conditions $u(x_0), u'(x_0), \dots, u^{(n-1)}(x_0)$.

There is *no such general theorem* about systems of first order PDE's. If there were, the study of PDE's would certainly be much simpler, because, as we will now point out, the Cauchy problem for any PDE can be reduced to the Cauchy problem for a system of first order PDE's. Because of the considerations in the previous section, we will assume that the partials of u with respect to one of the variables, which we will call y , are explicitly expressed in terms of the partials with respect to the other variables, which we will call x_1, \dots, x_n . Thus we consider the equation

$$\frac{\partial^k u}{\partial y^k}(x_1, \dots, x_n, y) = f\left(x, y, u(x, y), \dots, \frac{\partial^l u}{\partial x_1^{i_1} \dots \partial x_n^{i_n} \partial y^j}, \dots\right);$$

the partial derivatives appearing on the right are all of order $l \leq k$, and the order j with respect to y is $\leq k - 1$.

In the now-standard trick* for this reduction, we might as well allow our equation to be a system itself; in other words, u can be a vector-valued function. If we let $u_i = \partial u / \partial x_i$ and $v = \partial u / \partial y$, then we have

$$\begin{aligned} \frac{\partial^{k-1} u}{\partial y^{k-1}} &= \frac{\partial^{k-2} \frac{\partial u}{\partial y}}{\partial y^{k-2}} = \frac{\partial^{k-2} v}{\partial y^{k-2}} \\ \frac{\partial^{k-1} u_i}{\partial y^{k-1}} &= \frac{\partial^{k-1} \frac{\partial u}{\partial x_i}}{\partial y^{k-1}} \\ &= \frac{\partial}{\partial x_i} \frac{\partial^{k-1} u}{\partial y^{k-1}} = \frac{\partial}{\partial x_i} \frac{\partial^{k-2} \frac{\partial u}{\partial y}}{\partial y^{k-2}} = \frac{\partial}{\partial x_i} \frac{\partial^{k-2} v}{\partial y^{k-2}} \\ \frac{\partial^{k-1} v}{\partial y^{k-1}} &= f(x, y, u(x, y), \dots), \end{aligned}$$

with obvious initial conditions for u, u_1, \dots, u_n, v . Notice that the order of all partial derivatives of u with respect to y on the right is now $\leq k - 2$, so that this is a system of order $\leq k - 1$; conversely, this Cauchy problem for this system of (vector-valued) functions u, u_1, \dots, u_n, v gives us a solution for our original Cauchy problem. Thus, by repeating this process enough times we finally obtain a system of first order PDE's.

As a simple specific example, where the notation will be less abstract, consider the Cauchy problem for the equation

$$u_{yy} = f(x, y, u, u_x, u_y, u_{xx}, u_{xy}).$$

We want to introduce u_1 for $\partial u / \partial x$ and v for $\partial u / \partial y$; for convenience, we will simply use α for $\partial u / \partial x$. Then our Cauchy problem is equivalent to a Cauchy problem for the first order system

$$\begin{aligned} (*) \quad u_y &= v \\ \alpha_y &= v_x \\ v_y &= f(x, y, u, \alpha, v, \alpha_x, v_x). \end{aligned}$$

for (u, α, v) .

* From mimeographed notes, J. F. Trèves, *Ovsjannikov theorem and hyperdifferential operators*, Notas Mat. **46** (1968); see L. Nirenberg, *An abstract form of the nonlinear Cauchy-Kowalewski theorem*, J. Differential Geometry, **6** (1972), 561-576.

Yet one further simplification is possible: we can reduce any first order system

$$u^i_y = F(x, y, \dots, u^j, \dots, u^j_x, \dots)$$

to a *quasi-linear* system. To do this, we simply introduce new functions p^j (representing $\partial u/\partial x_j$) and consider the system

$$\begin{aligned} u^i_y &= F^i(x, y, \dots, u^j, \dots, p^j, \dots) \\ p^i_y &= F^i_x + \sum_j F^i_{u_j} \cdot u^j_x + \sum_j F^i_{p^j} \cdot p^j_x. \end{aligned}$$

For the system (*), instead of using the p^j notation, let us use*

$$\begin{aligned} p &\text{ for } u_x \\ r &\text{ for } \alpha_x \\ s &\text{ for } v_x. \end{aligned}$$

Then the Cauchy problem for (*) is equivalent to a Cauchy problem for the system

$$\begin{aligned} (**) \quad & u_y = v \\ & \alpha_y = s \\ & v_y = f(x, y, u, \alpha, v, r, s) \\ & p_y = v_y \\ & r_y = s_x \\ & s_y = f_x + f_u \cdot p + f_\alpha \cdot r + f_v \cdot s + f_r \cdot r_x + f_s \cdot s_x. \end{aligned}$$

4. THE CAUCHY-KOWALEWSKI THEOREM

In this section we will consider the most general system of first order quasi-linear equations in the variables x_1, \dots, x_n, y , where the partials with respect to y are expressed in terms of the partials with respect to x_1, \dots, x_n . We thus have N equations for N unknown functions u_1, \dots, u_N :

$$\begin{aligned} \frac{\partial u_\alpha}{\partial y} &= \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1, \dots, x_n, y, u_1, \dots, u_N) \frac{\partial u_\beta}{\partial x_i} \\ &\quad + B_\alpha(x_1, \dots, x_n, y, u_1, \dots, u_N). \end{aligned}$$

* For second order equations in 2 variables, the symbols p, q, r, s, t are customarily used for $u_x, u_y, u_{xx}, u_{xy}, u_{yy}$.

[Notice that the symbol $A_{\alpha i}^{\beta}(x_1, \dots, x_n, y, u_1, \dots, u_N)$ is really an abbreviation for

$$A_{\alpha i}^{\beta}(x_1, \dots, x_n, y, u_1(x_1, \dots, x_n, y), \dots, u_N(x_1, \dots, x_n, y)),$$

and similarly for B_{α} .] We will prove that this system of equations has a solution u_1, \dots, u_N with given initial conditions

$$u_{\alpha}(x_1, \dots, x_n, 0) = \xi_{\alpha}(x_1, \dots, x_n).$$

The hitch is that we will have to assume that both the coefficients $A_{\alpha i}^{\beta}, B_{\alpha}$ and the initial conditions ξ_{α} are *real analytic*. Recall that a function $f: U \subset \mathbb{R}^m \rightarrow \mathbb{R}$ is real analytic if it can be expressed as a convergent sum

$$\sum_{\sigma_1, \dots, \sigma_m=0}^{\infty} c_{\sigma_1 \dots \sigma_m} (x_1 - a_1)^{\sigma_1} \cdots (x_m - a_m)^{\sigma_m}$$

in a neighborhood of each point (a_1, \dots, a_m) in its domain. We will also write this in the abbreviated form

$$\sum_{\sigma} c_{\sigma} (x - a)^{\sigma}.$$

4. **THEOREM (THE CAUCHY-KOWALEWSKI THEOREM).** Let ξ_{α} ($\alpha = 1, \dots, N$) be analytic functions in a neighborhood of (a_1, \dots, a_n) in \mathbb{R}^n , set $b_{\alpha} = \xi_{\alpha}(a_1, \dots, a_n)$, and let $A_{\alpha i}^{\beta}$ and B_{α} ($\alpha, \beta = 1, \dots, N; i = 1, \dots, n$) be analytic functions in a neighborhood of $(a_1, \dots, a_n, 0, b_1, \dots, b_N)$ in \mathbb{R}^{N+n+1} . Then there are unique analytic functions u_1, \dots, u_N in a neighborhood of (a_1, \dots, a_n) in \mathbb{R}^n satisfying

$$\begin{aligned} \frac{\partial u_{\alpha}}{\partial y} = & \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^{\beta}(x_1, \dots, x_n, y, u_1, \dots, u_N) \frac{\partial u_{\beta}}{\partial x_i} \\ & + B_{\alpha}(x_1, \dots, x_n, y, u_1, \dots, u_N) \end{aligned}$$

with the initial conditions

$$u_{\alpha}(x_1, \dots, x_n, 0) = \xi_{\alpha}(x_1, \dots, x_n).$$

PROOF. We first make three slight simplifications.

(1) We can assume that all $a_i = 0$. For if we define

$$v_{\alpha}(x_1, \dots, x_n, y) = u_{\alpha}(x_1 + a_1, \dots, x_n + a_n, y),$$

then our equations and initial conditions are equivalent to equations of the same sort for v_α ,

$$\begin{aligned} \frac{\partial v_\alpha}{\partial y} = & \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1 + a_1, \dots, x_n + a_n, y, v_1, \dots, v_N) \frac{\partial v_\beta}{\partial x_i} \\ & + B_\alpha(x_1 + a_1, \dots, x_n + a_n, y, v_1, \dots, v_N), \end{aligned}$$

with the initial conditions

$$v_\alpha(x_1, \dots, x_n, 0) = \xi_\alpha(x_1 + a_1, \dots, x_n + a_n).$$

The functions $\bar{\xi}_\alpha(x_1, \dots, x_n) = \xi_\alpha(x_1 + a_1, \dots, x_n + a_n)$ are analytic in a neighborhood of 0 in \mathbb{R}^n , and the coefficients of the new equation are analytic in a neighborhood of $(0, \dots, 0, 0, \bar{\xi}_1(0, \dots, 0), \dots, \bar{\xi}_N(0, \dots, 0))$ in \mathbb{R}^{N+n+1} .

(2) We can further assume that the ξ_α are all 0. For if we now define

$$v_\alpha(x_1, \dots, x_n, y) = u_\alpha(x_1, \dots, x_n, y) - \xi_\alpha(x_1, \dots, x_n),$$

and for abbreviation set

$$C_k = v_k + \xi_k(x_1, \dots, x_n) \quad k = 1, \dots, N,$$

then our equations and initial conditions are equivalent to equations of the same sort for the v_α ,

$$\begin{aligned} \frac{\partial v_\alpha}{\partial y} = & \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1, \dots, x_n, y, C_1, \dots, C_N) \frac{\partial v_\beta}{\partial x_i} \\ & + \left[\sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1, \dots, x_n, y, C_1, \dots, C_N) \cdot \frac{\partial \xi_\beta}{\partial x_i}(x_1, \dots, x_n) \right. \\ & \left. + B_\alpha(x_1, \dots, x_n, y, C_1, \dots, C_N) \right], \end{aligned}$$

with the initial conditions

$$v_\alpha(x_1, \dots, x_n, 0) = 0.$$

Notice that the coefficients of the new equation are analytic in a neighborhood of 0 in \mathbb{R}^{N+n+1} .

(3) We can assume finally that the $A_{\alpha i}^\beta$ and the B_α do not depend on y . For we can consider the equations in $N + 1$ unknowns η, u_1, \dots, u_N ,

$$\begin{aligned} \frac{\partial \eta}{\partial y} &= 1 \\ \frac{\partial u_\alpha}{\partial y} &= \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1, \dots, x_n, \eta, u_1, \dots, u_N) \frac{\partial u_\beta}{\partial x_i} \\ &\quad + B_\alpha(x_1, \dots, x_n, \eta, u_1, \dots, u_N), \end{aligned}$$

with initial conditions

$$\begin{aligned} \eta(x_1, \dots, x_n, 0) &= 0 \\ u_\alpha(x_1, \dots, x_n, 0) &= 0. \end{aligned}$$

To sum up, we can consider equations

$$\begin{aligned} (1) \quad \frac{\partial u_\alpha}{\partial y} &= \sum_{\beta=1}^N \sum_{i=1}^n A_{\alpha i}^\beta(x_1, \dots, x_n, u_1, \dots, u_N) \frac{\partial u_\beta}{\partial x_i} \\ &\quad + B_\alpha(x_1, \dots, x_n, u_1, \dots, u_N) \end{aligned}$$

with initial conditions

$$(2) \quad u_\alpha(x_1, \dots, x_n, 0) = 0;$$

the functions $A_{\alpha i}^\beta$ and B_α are analytic in a neighborhood of 0 in \mathbb{R}^{N+n} , and we are looking for a solution u in a neighborhood of 0. We expand the analytic functions $A_{\alpha i}^\beta$ and B_α around 0 as

$$\begin{aligned} (3) \quad A_{\alpha i}^\beta(x_1, \dots, x_n, z_1, \dots, z_N) &= \sum a_{\alpha i; \sigma_1, \dots, \sigma_n, \tau_1, \dots, \tau_N}^\beta x_1^{\sigma_1} \dots x_n^{\sigma_n} z_1^{\tau_1} \dots z_N^{\tau_N} \\ &= \sum_{\sigma, \tau} a_{\alpha i; \sigma, \tau}^\beta x^\sigma z^\tau \end{aligned}$$

$$\begin{aligned} (4) \quad B_\alpha(x_1, \dots, x_n, z_1, \dots, z_N) &= \sum b_{\alpha; \sigma_1, \dots, \sigma_n, \tau_1, \dots, \tau_N} x_1^{\sigma_1} \dots x_n^{\sigma_n} z_1^{\tau_1} \dots z_N^{\tau_N} \\ &= \sum_{\sigma, \tau} b_{\alpha; \sigma, \tau} x^\sigma z^\tau. \end{aligned}$$

We claim, first of all, that there is at most one analytic solution

$$\begin{aligned} (5) \quad u_\alpha(x_1, \dots, x_n, y) &= \sum c_{\alpha; \sigma_1, \dots, \sigma_n, \rho} x_1^{\sigma_1} \dots x_n^{\sigma_n} y^\rho \\ &= \sum_{\sigma, \rho} c_{\alpha; \sigma, \rho} x^\sigma y^\rho \end{aligned}$$

of (1) and (2). We just have to show that the coefficients $c_{\alpha;\sigma,\rho}$ are completely determined by those of $A_{\alpha i}^\beta$ and B_α . From the particular way they are determined, we will then be able to show that the resulting series (5) converges, thereby also proving existence.

For a given n -tuple $\sigma = (\sigma_1, \dots, \sigma_n)$, let $\sigma + \delta_i$ be the n -tuple

$$\sigma + \delta_i = (\sigma_1, \dots, \sigma_i + 1, \dots, \sigma_n).$$

Then if the u_α are given by (5), we can write

$$(6) \quad \begin{cases} \frac{\partial u_\beta}{\partial x_i}(x_1, \dots, x_n, y) = \sum_{\sigma, \rho} (\sigma_i + 1) c_{\beta; \sigma + \delta_i, \rho} x^\sigma y^\rho \\ \frac{\partial u_\alpha}{\partial y}(x_1, \dots, x_n, y) = \sum_{\sigma, \rho} (\rho + 1) c_{\alpha; \sigma, \rho+1} x^\sigma y^\rho. \end{cases}$$

So if the u_α in (5) satisfy (1), then

$$(7) \quad \begin{cases} \sum_{\sigma, \rho} (\rho + 1) c_{\alpha; \sigma, \rho+1} x^\sigma y^\rho \\ = \sum_{\beta=1}^N \sum_{i=1}^n \left\{ \sum_{\sigma, \tau} a_{\alpha i; \sigma, \tau}^\beta x^\sigma \left(\sum_{\sigma, \rho} c_{1; \sigma, \rho} x^\sigma y^\rho \right)^{\tau_1} \cdots \left(\sum_{\sigma, \rho} c_{N; \sigma, \rho} x^\sigma y^\rho \right)^{\tau_N} \right\} \\ \quad \cdot \sum_{\sigma, \rho} (\sigma_i + 1) c_{\beta; \sigma + \delta_i, \rho} x^\sigma y^\rho \\ + \sum_{\sigma, \tau} b_{\alpha; \sigma, \tau} x^\sigma \left(\sum_{\sigma, \rho} c_{1; \sigma, \rho} x^\sigma y^\rho \right)^{\tau_1} \cdots \left(\sum_{\sigma, \rho} c_{N; \sigma, \rho} x^\sigma y^\rho \right)^{\tau_N}. \end{cases}$$

Now there is no need to become unduly frightened by this expression. After we expand everything out, we will have an expression of the form

$$(8) \quad \begin{aligned} \sum_{\sigma, \rho} (\rho + 1) c_{\alpha; \sigma, \rho+1} x^\sigma y^\rho &= \sum_{\alpha, \sigma, \rho} P_{\alpha, \sigma, \rho} (a_{\xi j; \mu, v}^\eta, b_{\xi; \mu, v}, c_{\xi; \mu, v}) \\ &= \sum_{\alpha, \sigma, \rho} P_{\alpha, \sigma, \rho} (a, b, c) \quad \text{for short,} \end{aligned}$$

where each $P_{\alpha, \sigma, \rho}$ is a polynomial in certain of the $a_{\xi j; \mu, v}^\eta$, $b_{\xi; \mu, v}$, and $c_{\xi; \mu, v}$. Just which of these appear as arguments of $P_{\alpha, \sigma, \rho}$ depends on (σ, ρ) ; the only important thing for us to note is that

$$(A) \quad P_{\alpha, \sigma, \rho} \text{ depends only on those } c_{\xi; \mu, v} \text{ with } v \leq \rho.$$

Notice that all the information in equation (1) enters into (8) as the *arguments* of the polynomials $P_{\alpha,\sigma,\rho}$. These polynomials themselves *do not depend* on the $A_{\alpha i}^\beta$ or B_α , or on the $c_{\xi;\mu,\nu}$:

(B) The polynomials $P_{\alpha,\sigma,\rho}$ are “universal polynomials” depending only on N and n .

Finally, we note that

(C) The coefficients of $P_{\alpha,\sigma,\rho}$ are *non-negative* integers.

Now if (8) is to hold for all sufficiently small (x_1, \dots, x_n, y) , then we must have

$$(9) \quad (\rho + 1)c_{\alpha;\sigma,\rho+1} = P_{\alpha,\sigma,\rho}(a, b, c).$$

Together with the initial condition (2), which gives us $c_{\alpha;\sigma,0} = 0$, we can now calculate all $c_{\alpha;\sigma,\rho}$ recursively from (9), since (A) shows us that the right side involves only $c_{\xi;\mu,\nu}$ with $\nu \leq \rho$. This proves uniqueness.

To prove existence, we must show that the series (5) converges, when the $c_{\alpha;\sigma,\rho}$ are computed from (9). We will show absolute convergence for sufficiently small (x_1, \dots, x_n, y) . This is done by the following trick, called the **method of majorants**. Consider another set of equations

$$(1') \quad \begin{aligned} \frac{\partial u_\alpha}{\partial y} = & \sum_{\beta=1}^N \sum_{i=1}^n \bar{A}_{\alpha i}^\beta(x_1, \dots, x_n, u_1, \dots, u_N) \frac{\partial u_\beta}{\partial x_i} \\ & + \bar{B}_\alpha(x_1, \dots, x_n, u_1, \dots, u_N) \end{aligned}$$

which “majorizes” (1), that is, which satisfies

$$(10) \quad \begin{cases} |a_{\xi j;\mu,\nu}^\eta| \leq \bar{a}_{\xi j;\mu,\nu}^\eta \\ |b_{\xi;\mu,\nu}| \leq \bar{b}_{\xi;\mu,\nu} \end{cases} \quad \text{for all } \xi, \eta, j, \mu, \nu.$$

Suppose that equation (1'), with the same initial condition (2), has an analytic solution

$$(5') \quad u_\alpha(x_1, \dots, x_n, y) = \sum_{\sigma, \rho} \bar{c}_{\alpha; \sigma, \rho} x^\sigma y^\rho.$$

Then we must have

$$(8') \quad (\rho + 1) \bar{c}_{\alpha; \sigma, \rho+1} = P_{\alpha, \sigma, \rho}(\bar{a}, \bar{b}, \bar{c}),$$

where, by (B), the $P_{\alpha, \sigma, \rho}$ are the *same* universal polynomials as in (8). We claim that we can then conclude that

$$(11) \quad |c_{\alpha; \sigma, \rho}| \leq \bar{c}_{\alpha; \sigma, \rho}.$$

The proof is by induction. It is clear for $\rho = 0$. Now assume it is true for ρ . Then

$$\begin{aligned} |c_{\alpha; \sigma, \rho+1}| &= \frac{1}{\rho + 1} |P_{\alpha, \sigma, \rho}(a, b, c)| && \text{by (8)} \\ &\leq \frac{1}{\rho + 1} P_{\alpha, \sigma, \rho}(|a|, |b|, |c|) && \text{by (C)} \\ &\leq \frac{1}{\rho + 1} P_{\alpha, \sigma, \rho}(\bar{a}, \bar{b}, \bar{c}) && \text{by (C) and (10)} \\ &= \bar{c}_{\alpha; \sigma, \rho+1} && \text{by (8')}. \end{aligned}$$

This completes the induction proof. But now the convergence of (5'), together with (11), proves the absolute convergence of (5). So the proof of the theorem will be complete once we show that some majorant of equation (1) has an analytic solution. This will be comparatively easy.

For some $r > 0$, the power series (3) converges for the point (x, z) with all $x_j = z_\xi = r$. Then the terms

$$a_{\alpha i; \sigma, \tau}^\beta r^{\sigma_1 + \dots + \tau_N}$$

in this infinite sum must approach 0, and consequently are surely bounded: there is some M with

$$|a_{\alpha i; \sigma, \tau}^\beta| \leq \frac{M}{r^{\sigma_1 + \dots + \tau_N}}$$

for all σ, τ . For r small enough and M large enough, this equation holds for each $\alpha, \beta \leq N$ and $i \leq n$. Similarly, we can assume that we have the same

estimate for each $|b_{\alpha;\sigma,\tau}|$. We thus also have the weaker estimates

$$|a_{\alpha i;\sigma,\tau}^\beta|, |b_{\alpha;\sigma,\tau}| \leq \frac{M}{r^{\sigma_1+\dots+\tau_N}} \cdot \frac{(\sigma_1 + \dots + \tau_N)!}{\sigma_1! \dots \tau_N!}.$$

So if we take $\bar{a}_{\alpha i;\sigma,\tau}^\beta$ and $\bar{b}_{\alpha;\sigma,\tau}$ to be the expression on the right side of this inequality, then equation (1') majorizes equation (1), and we just have to show that the solution of (1') with the initial condition (2) is analytic. Since

$$\begin{aligned} \sum_{\sigma,\tau} \bar{a}_{\alpha i;\sigma,\tau}^\beta x^\sigma z^\tau &= \sum_{\sigma,\tau} \bar{b}_{\alpha;\sigma,\tau} x^\sigma z^\tau \\ &= M \sum_{\sigma,\tau} \left(\frac{x}{r}\right)^\sigma \left(\frac{z}{r}\right)^\tau \frac{(\sigma_1 + \dots + \tau_N)!}{\sigma_1! \dots \tau_N!} \\ &= \frac{M}{1 - \frac{(x_1 + \dots + z_N)}{r}}, \end{aligned}$$

we are dealing with the equations

$$\begin{cases} \frac{\partial u_\alpha}{\partial y} = \frac{M}{1 - \frac{(x_1 + \dots + u_N)}{r}} \left(\sum_{\beta=1}^N \sum_{i=1}^n \frac{\partial u_\beta}{\partial x_i} + 1 \right) \\ u_\alpha(x_1, \dots, x_n, 0) = 0. \end{cases}$$

Since the equations for the different u_α are all the same, and since the x_i enter only in the combination $X = x_1 + \dots + x_n$, it seems reasonable to look for a solution with all

$$u_\alpha(x_1, \dots, x_n, y) = U(x_1 + \dots + x_n, y) = U(X, y).$$

This gives us the single equation

$$\begin{cases} \frac{\partial U}{\partial y} = \frac{M}{1 - \frac{X + NU}{r}} \left(Nn \frac{\partial U}{\partial x} + 1 \right) \\ U(X, 0) = 0. \end{cases}$$

This is a first order equation, and hence one which, in theory, we can deal with.

The simplest thing, at the moment, is just to check that

$$U(X, y) = \frac{r - X}{(n + 1)N} - \frac{\sqrt{(r - X)^2 - 2N(n + 1)Mr y}}{N(n + 1)}$$

is the desired solution, and that it is analytic in a neighborhood of 0. ♦

Although the Cauchy-Kowalewski Theorem is the most general result in the theory of PDE's, its usefulness is greatly restricted by the fact that both the *coefficients* and the *initial conditions* must be real analytic. We would naturally like to know whether these hypotheses are somehow dictated by the very nature of the problem, or whether they represent merely a defect in our method of proof. [One source of difficulty may be the fact that in one respect the theorem proves too much, since it is formulated for an *arbitrary* system of first order quasi-linear equations. Although it would be nice to solve any such system, this problem does not bear directly on the problem of solving a single higher order PDE, because only very special sorts of systems of first order equations are derived from higher order equations; given an arbitrary system of first order equations with initial conditions, we generally cannot find a single higher order equation with initial conditions that is equivalent to it.]

The necessity of having *analytic coefficients* in the Cauchy-Kowalewski Theorem is demonstrated by the famous example of Hans Lewy [1],

$$\begin{cases} \frac{\partial u_1}{\partial x_1} = \frac{\partial u_2}{\partial x_2} - 2x_2 \frac{\partial u_1}{\partial x_3} - 2x_1 \frac{\partial u_2}{\partial x_3} - f'(x_3) \\ \frac{\partial u_2}{\partial x_1} = -\frac{\partial u_1}{\partial x_2} + 2x_1 \frac{\partial u_1}{\partial x_3} - 2x_2 \frac{\partial u_2}{\partial x_3}. \end{cases}$$

If f is C^∞ but not analytic, then this system has no solutions at all (let alone solutions with arbitrary initial conditions). By the way, this system can be considered as the following single equation for the complex-valued function $u = u_1 + iu_2$:

$$-\frac{\partial u}{\partial x_1} - i \frac{\partial u}{\partial x_2} + 2i(x_1 + ix_2) \frac{\partial u}{\partial x_3} = f'(x_3).$$

There are also cases where *analytic initial conditions* are necessary, for we will soon see that there are simple PDE's, with analytic coefficients, that cannot have solutions with given initial conditions unless these conditions are analytic. So in a certain respect the Cauchy-Kowalewski theorem gives the best possible result in these cases. On the other hand, it turns out that the Cauchy problem isn't even the one which we want to pose for these equations. Moreover, there is a wide class of equations where the Cauchy problem is a natural one, but where analyticity is much too severe a restriction.

5. CLASSIFICATION OF SECOND ORDER PDE's

In our first forage into the uncharted lands of higher order PDE's, it is natural that we first restrict our attention to those of second order:

$$F\left(x_1, \dots, x_n, u(x_1, \dots, x_n), \frac{\partial u}{\partial x_1}(x_1, \dots, x_n), \dots, \frac{\partial^2 u}{\partial x_n^2}(x_1, \dots, x_n)\right) = 0.$$

As a matter of fact, we will never get anywhere beyond this. Moreover, we will deal almost exclusively with equations in only two variables,

$$F\left(x, y, u(x, y), \frac{\partial u}{\partial x}(x, y), \frac{\partial u}{\partial y}(x, y), \frac{\partial^2 u}{\partial x^2}(x, y), \frac{\partial^2 u}{\partial x \partial y}(x, y), \frac{\partial^2 u}{\partial y^2}(x, y)\right) = 0,$$

or in abbreviated form

$$F(x, y, u, p, q, r, s, t) = 0.$$

Even nowadays there are certain phenomena about second order PDE's which are much more completely understood in the two variable case than in higher dimensions, but the particular results that we are after can all be handled in a uniform way that works in all dimensions. However, a whole book would be required in order to reach them. So we will instead use quite classical methods to analyze second order PDE's in just 2 variables. Fortunately, the 2 variable case happens to be just the one we are interested in.

We begin by singling out the *semi-linear* equations

$$\begin{aligned} \text{(I)} \quad 0 &= a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + f(x, y, u, u_x, u_y) \\ &= L(u) + f. \end{aligned}$$

For such an equation, $L(u)$ is called the “principal part”, and we will often denote the remaining part, involving only lower order derivatives, by “...”. There is a classification for these equations which is closely related to the classification of algebraic equations of the form

$$z = ax^2 + 2bxy + cy^2 = \left\langle (x, y), (x, y) \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \right\rangle,$$

with a, b, c not all 0. We briefly remind the reader how this classification goes (it is essentially the same as the classification of points on a surface in Chapter 2). We choose two orthonormal eigenvectors $X_1, X_2 \in \mathbb{R}^2$ for the symmetric matrix $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$, with corresponding eigenvalues λ_1, λ_2 . Then

$$\begin{aligned} \left\langle \phi X_1 + \psi X_2, (\phi X_1 + \psi X_2) \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \right\rangle &= \langle \phi X_1 + \psi X_2, \phi \lambda_1 X_1 + \psi \lambda_2 X_2 \rangle \\ &= \lambda_1 \phi^2 + \lambda_2 \psi^2. \end{aligned}$$

So if we use ϕ and ψ as new coordinates for \mathbb{R}^2 , our equation becomes

$$z = \lambda_1 \phi^2 + \lambda_2 \psi^2.$$

We can express this statement a little more precisely in terms of the linear transformation $S = (\phi, \psi): \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by

$$S(X_1) = (1, 0), \quad S(X_2) = (0, 1).$$

Since

$$S(x, y) = (\phi(x, y), \psi(x, y)) = \phi(x, y)S(X_1) + \psi(x, y)S(X_2),$$

the functions ϕ and ψ are just the coordinates of (x, y) with respect to X_1 and X_2 :

$$(x, y) = \phi(x, y)X_1 + \psi(x, y)X_2.$$

So we obtain

$$\begin{aligned} \text{(II)} \quad & ax^2 + 2bxy + cy^2 \\ &= \left\langle (x, y), (x, y) \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \right\rangle \\ &= \left\langle \phi(x, y)X_1 + \psi(x, y)X_2, (\phi(x, y)X_1 + \psi(x, y)X_2) \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \right\rangle \\ &= \langle \phi(x, y)X_1 + \psi(x, y)X_2, \phi(x, y)\lambda_1 X_1 + \psi(x, y)\lambda_2 X_2 \rangle \\ &= \lambda_1 [\phi(x, y)]^2 + \lambda_2 [\psi(x, y)]^2. \end{aligned}$$

These algebraic manipulations are often expressed slightly differently. We can write equation (II) as

$$\begin{aligned} (x, y) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= (\phi(x, y), \psi(x, y)) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \phi(x, y) \\ \psi(x, y) \end{pmatrix} \\ &= S(x, y) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} [S(x, y)]^t, \end{aligned}$$

where t denotes the transpose. If Q is the matrix of the linear transformation S , then we can write

$$\begin{aligned} (x, y) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= [(x, y) \cdot Q] \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} [(x, y) \cdot Q]^t \\ &= (x, y) \left[Q \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} Q^t \right] \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{for all } (x, y), \end{aligned}$$

which implies that

$$(III) \quad \begin{pmatrix} a & b \\ b & c \end{pmatrix} = Q \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} Q^t.$$

It is not hard to see exactly what the matrix Q is. Since S^{-1} takes $(1, 0)$ to X_1 and $(0, 1)$ to X_2 , its matrix has X_1 and X_2 as its two columns,

$$Q^{-1} = (X_1^t, X_2^t).$$

Moreover, since X_1 and X_2 are orthonormal, we have $Q^{-1}(Q^{-1})^t = I$. So

$$Q = (Q^{-1})^t = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Since $Q^{-1} = Q^t$, we can also write (III) as

$$(IV) \quad \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = Q^t \begin{pmatrix} a & b \\ b & c \end{pmatrix} Q.$$

The reduction (II) [or its equivalent (IV)] shows that the equations $z = ax^2 + 2bxy + cy^2$ fall into three classes:

Elliptic Case: $ac - b^2 > 0$; equivalently, λ_1 and λ_2 have the same sign

Hyperbolic Case: $ac - b^2 < 0$; equivalently, λ_1 and λ_2 have opposite signs

Parabolic Case: $ac - b^2 = 0$; $\lambda_1 = 0$ or $\lambda_2 = 0$.

We introduce a similar classification for semi-linear PDE's

$$(I) \quad a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + \cdots = 0.$$

If a_0, b_0, c_0 are the values of the functions a, b, c at (x_0, y_0) , then we say that equation (I) is

elliptic at (x_0, y_0) if $a_0c_0 - b_0^2 > 0$

hyperbolic at (x_0, y_0) if $a_0c_0 - b_0^2 < 0$

parabolic at (x_0, y_0) if $a_0c_0 - b_0^2 = 0$

(but not all of a_0, b_0, c_0 are 0).

Naturally we say that equation (I) is **elliptic** in an open set U if it is elliptic at each point $(x, y) \in U$, etc. The simplest examples of equations of these three types are the “normal forms”

$$(E) \quad u_{xx} + u_{yy} + \cdots = 0$$

$$(H) \quad u_{xx} - u_{yy} + \cdots = 0$$

$$(P) \quad u_{xx} + \cdots = 0.$$

As always, “ \dots ” denotes terms which do not involve any second derivatives. There is in addition an alternative normal form in the hyperbolic case,

$$(H') \quad u_{xy} + \cdots = 0$$

(corresponding to the possibility of writing the equation for a hyperbola in the form $xy = 1$).

We would like to see if equation (I) can be reduced to a normal form by writing it in terms of a function v defined by

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y));$$

here (ϕ, ψ) is supposed to be a differentiable map from \mathbb{R}^2 to \mathbb{R}^2 with differentiable inverse. Denoting a typical point in the domain of v by (ξ, η) , and the partials of v by v_ξ and v_η , we compute that

$$u_x = v_\xi \phi_x + v_\eta \psi_x, \quad u_y = v_\xi \phi_y + v_\eta \psi_y$$

and then that

$$(VI) \quad \begin{cases} u_{xx} = v_{\xi\xi} \phi_x^2 + 2v_{\xi\eta} \phi_x \psi_x + v_{\eta\eta} \psi_x^2 + \cdots \\ u_{xy} = v_{\xi\xi} \phi_x \phi_y + v_{\xi\eta} (\phi_x \psi_y + \phi_y \psi_x) + v_{\eta\eta} \psi_x \psi_y + \cdots \\ u_{yy} = v_{\xi\xi} \phi_y^2 + 2v_{\xi\eta} \phi_y \psi_y + v_{\eta\eta} \psi_y^2 + \cdots \end{cases}$$

where “ \dots ” again denotes terms which do not involve any second derivatives. [Naturally, if the derivatives of u are evaluated at (x, y) , then the derivatives of ϕ and ψ are evaluated at (x, y) , while those for v are evaluated at $(\phi(x, y), \psi(x, y))$.] From this we easily see that

$$(VII) \quad \begin{cases} au_{xx} + 2bu_{xy} + cu_{yy} = \alpha v_{\xi\xi} + 2\beta v_{\xi\eta} + \gamma v_{\eta\eta} + \cdots \\ \text{where} \\ \alpha = a\phi_x^2 + 2b\phi_x\phi_y + c\phi_y^2 \\ \beta = a\phi_x\psi_x + b(\phi_x\psi_y + \phi_y\psi_x) + c\phi_y\psi_y \\ \gamma = a\psi_x^2 + 2b\psi_x\psi_y + c\psi_y^2 \\ \text{or equivalently} \\ \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} = \begin{pmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \phi_x & \psi_x \\ \phi_y & \psi_y \end{pmatrix}. \end{cases}$$

Notice that the last part of (VII) shows that

$$(VIII) \quad \alpha\gamma - \beta^2 = (ac - b^2)(\phi_x\psi_y - \phi_y\psi_x)^2;$$

therefore the type of the equation for v is always the same as the type for u .

In one case, purely algebraic manipulations will reduce our equation to normal form:

5. PROPOSITION. Suppose that the equation

$$(I) \quad au_{xx} + bu_{xy} + cu_{yy} + \cdots = 0$$

has *constant* coefficients a, b, c in the principal part. Then there is a non-singular linear transformation $(\phi, \psi): \mathbb{R}^2 \rightarrow \mathbb{R}^2$ having the property that if v is defined by

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y)),$$

then u satisfies (I) if and only if v satisfies a certain equation of the form (E), (H), or (P). In the hyperbolic case, we can also find (ϕ, ψ) so that u satisfies (I) if and only if v satisfies a certain equation of the form (H').

PROOF. Equation (IV) shows that we can choose a constant matrix $Q = \begin{pmatrix} \phi_x & \psi_x \\ \phi_y & \psi_y \end{pmatrix}$ so that

$$\begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{in equation (VII).}$$

Since $\det Q \neq 0$, the linear transformation (ϕ, ψ) is non-singular. If we define v by (V), then (VII) shows that equation (I) for u is equivalent to the equation

$$(I) \quad \lambda_1 v_{\xi\xi} + \lambda_2 v_{\eta\eta} + \cdots = 0$$

for v . If we make a further change of coordinates by defining

$$\tilde{v}(\rho, \sigma) = v(r\rho, s\sigma) \quad r, s \text{ constants,}$$

then

$$\tilde{v}_{\rho\rho} = r^2 v_{\xi\xi} \quad \text{and} \quad \tilde{v}_{\sigma\sigma} = s^2 v_{\eta\eta}.$$

So we can also arrange for λ_1 and λ_2 to be ± 1 in (I). Then we have equations equivalent to (E), (H), or (P) [we may have to interchange the names of ξ and η].

The form (H') is obtained by analogy with the fact that the equation $x^2 - y^2 = 1$ becomes $4\bar{x}\bar{y} = 1$ when we perform the substitution $\bar{x} = \frac{1}{2}(x + y)$, $\bar{y} = \frac{1}{2}(x - y)$. We define

$$(2) \quad w(\rho, \sigma) = v\left(\frac{\rho + \sigma}{2}, \frac{\rho - \sigma}{2}\right).$$

Then

$$w_\rho = \frac{1}{2}v_\xi + \frac{1}{2}v_\eta$$

$$w_{\rho\sigma} = \frac{1}{2}\left(\frac{1}{2}v_{\xi\xi} - \frac{1}{2}v_{\xi\eta}\right) + \frac{1}{2}\left(\frac{1}{2}v_{\eta\xi} - \frac{1}{2}v_{\eta\eta}\right) = \frac{1}{4}(v_{\xi\xi} - v_{\eta\eta}).$$

So an equation of the form (H) for v is equivalent to one of the form (H') for w . ♦

The same method that was used in this proof will clearly enable us to reduce the general semi-linear equation (I) to an equation which has the normal form *at one point* (x_0, y_0) . But to obtain the normal form in a whole neighborhood, we have to work much harder. We consider the elliptic case first.

6. THEOREM. Suppose that the equation

$$(I) \quad a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + \cdots = 0$$

is elliptic in a neighborhood of (x_0, y_0) . Then there is a differentiable map (ϕ, ψ) from a neighborhood of (x_0, y_0) into \mathbb{R}^2 , with differentiable inverse, having the property that if v is defined by

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y)),$$

then u satisfies (I) if and only if v satisfies a certain equation in the normal form (E).

PROOF. It obviously suffices to find (ϕ, ψ) so that in equation (VII) we have $\alpha = \gamma$ and $\beta = 0$. So it suffices to find (ϕ, ψ) with

$$(I) \quad a\phi_x^2 + 2b\phi_x\phi_y + c\phi_y^2 = a\psi_x^2 + 2b\psi_x\psi_y + c\psi_y^2$$

$$(2) \quad a\phi_x\psi_x + b(\phi_x\psi_y + \phi_y\psi_x) + c\phi_y\psi_y = 0,$$

and $\phi_x\psi_y - \phi_y\psi_x \neq 0$ at (x_0, y_0) . This is precisely the problem of introducing isothermal coordinates for the metric $a dx \otimes dx + b[dx \otimes dy + dy \otimes dx] + c dy \otimes dy$, which we solved in Addendum I to Chapter 9. ♦

In our proof of the existence of isothermal coordinates, we showed that (1) and (2) are equivalent to the “Beltrami equations”

$$(a) \quad \phi_x = \frac{b\psi_x + c\psi_y}{\sqrt{ac - b^2}}, \quad \phi_y = -\frac{a\psi_x + b\psi_y}{\sqrt{ac - b^2}}.$$

Note that if (a) is to hold, then we must have

$$(b) \quad \frac{\partial}{\partial x} \left(\frac{a\psi_x + b\psi_y}{W} \right) + \frac{\partial}{\partial y} \left(\frac{b\psi_x + c\psi_y}{W} \right) = 0, \quad W = \sqrt{ac - b^2}.$$

Conversely, if (b) holds for some ϕ , then there is ψ satisfying (a); moreover, the Jacobian of (ϕ, ψ) is

$$\phi_x \psi_y - \phi_y \psi_x = -\frac{1}{W}(a\phi_x^2 + 2b\phi_x \phi_y + c\phi_y^2),$$

which is non-zero if $(\phi_x, \phi_y) \neq (0, 0)$. So solving (a) is equivalent to solving (b), which is itself elliptic [with the very same principal part as (I)]. Had we not already solved equation (a), we would be in the embarrassing position of needing to know that elliptic equations have solutions before we could reduce them to normal form.

In the hyperbolic case, the same line of reasoning leads us into precisely this difficulty, and thus requires results from section 7. However, there is also an elementary argument.

7. THEOREM. Suppose that the equation

$$(I) \quad a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + \cdots = 0$$

is hyperbolic at (x_0, y_0) . Then there is a differentiable map (ϕ, ψ) from a neighborhood of (x_0, y_0) into \mathbb{R}^2 , with differentiable inverse, having the property that if v is defined by

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y)),$$

then u satisfies (I) if and only if v satisfies a certain equation in the normal form (H). The same result holds for a certain equation in the normal form (H').

FIRST PROOF. We claim, first, that we can assume that $c_0 = c(x_0, y_0) \neq 0$. For suppose that $c_0 = 0$. Choose (ϕ, ψ) to be a linear transformation with matrix $\begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}$. Then at (x_0, y_0) the coefficients α, β, γ of the equation for v

are, by (VII),

$$\begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix} \begin{pmatrix} a_0 & b_0 \\ b_0 & 0 \end{pmatrix} \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a_0 & \lambda a_0 + b_0 \\ \lambda a_0 + b_0 & \lambda^2 a_0 + 2\lambda b_0 \end{pmatrix}.$$

Since we must have $a_0 \neq 0$ or $b_0 \neq 0$, we can certainly choose λ with $\lambda^2 a_0 + 2\lambda b_0 \neq 0$.

So we assume that $c(x_0, y_0) \neq 0$. To achieve the normal form (H), it now clearly suffices to choose (ϕ, ψ) so that in equation (VII) we have $\alpha = -\gamma$ and $\beta = 0$. Solving as before, we end up with the system of equations

$$\begin{aligned} \phi_x &= \frac{b\psi_x + c\psi_y}{\sqrt{b^2 - ac}} \\ \phi_y &= -\frac{a\psi_x + b\psi_y}{\sqrt{b^2 - ac}}, \end{aligned}$$

or equivalently

$$\begin{cases} \phi_y = -\frac{b}{c}\phi_x + \frac{\sqrt{b^2 - ac}}{c}\psi_x \\ \psi_y = \frac{\sqrt{b^2 - ac}}{c}\phi_x - \frac{b}{c}\psi_x. \end{cases}$$

Section 7 shows that we can solve this ("hyperbolic") system.

To obtain the normal form (H'), we start with v satisfying an equation of the normal form (H),

$$v_{\xi\xi} - v_{\eta\eta} + \cdots = 0,$$

and define

$$(2) \quad w(\rho, \sigma) = v\left(\frac{\rho + \sigma}{2}, \frac{\rho - \sigma}{2}\right).$$

As in the proof of Proposition 5, we find that w satisfies an equation of the normal form (H').

SECOND PROOF. We can instead try for the normal form (H') directly; the normal form (H) is then obtained by the same change of coordinates used in (2), which is equal to its own inverse, up to a factor of 2. So it suffices to find (ϕ, ψ) so that in equation (VII) we have $\alpha = \gamma = 0$; thus we need

$$(1) \quad \begin{cases} a\phi_x^2 + 2b\phi_x\phi_y + c\phi_y^2 = 0 \\ a\psi_x^2 + 2b\psi_x\psi_y + c\psi_y^2 = 0 \end{cases}$$

and

$$(2) \quad \phi_x \psi_y - \phi_y \psi_x \neq 0.$$

As in the previous proof, we can assume that $c(x_0, y_0) \neq 0$. If there is any hope of solving (1) and (2), we clearly must have $\phi_x, \psi_x \neq 0$. This suggests that we look at the equations

$$(1') \quad \begin{cases} a + 2b \left(\frac{\phi_y}{\phi_x} \right) + c \left(\frac{\phi_y}{\phi_x} \right)^2 = 0 \\ a + 2b \left(\frac{\psi_y}{\psi_x} \right) + c \left(\frac{\psi_y}{\psi_x} \right)^2 = 0 \end{cases}$$

$$(2') \quad \frac{\phi_y}{\phi_x} \neq \frac{\psi_y}{\psi_x}.$$

Clearly (1') and (2') imply (1) and (2). Now $ac - b^2 < 0$, so the equation

$$(3) \quad a(x, y) + 2b(x, y)\mu + c(x, y)\mu^2 = 0 \quad c(x, y) \neq 0$$

always has two *distinct, real* roots, $\mu_1(x, y)$ and $\mu_2(x, y)$, varying continuously with (x, y) . So we just have to find ϕ and ψ satisfying

$$(4) \quad \begin{cases} \phi_y - \mu_1 \phi_x = 0 & \phi_x \neq 0 \text{ at } (x_0, y_0) \\ \psi_y - \mu_2 \psi_x = 0 & \psi_x \neq 0 \text{ at } (x_0, y_0), \end{cases}$$

in order for (1') and (2') to hold. But the two equations in (4) are each linear first order PDE's, and the line $y = y_0$ is free, for any given initial conditions $\phi(x, y_0)$ and $\psi(x, y_0)$; in particular, we can assure that $\phi_y(x_0, y_0), \psi_y(x_0, y_0) \neq 0$. ♦

Finally, there is no problem in the parabolic case.

8. THEOREM. Suppose that the equation

$$(I) \quad a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + \cdots = 0$$

is parabolic in a neighborhood of (x_0, y_0) . Then there is a differentiable map (ϕ, ψ) from a neighborhood of (x_0, y_0) into \mathbb{R}^2 , with differentiable inverse, having the property that if v is defined by

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y)),$$

then u satisfies (I) if and only if v satisfies a certain equation in the normal form (P).

PROOF. It obviously suffices to find (ϕ, ψ) with $\phi_x \psi_y - \phi_y \psi_x \neq 0$ so that in equation (VII) we have $\gamma = 0$. For then β must be 0 by (VIII), while the last part of (VII) shows that α cannot be zero. We thus want

$$(I) \quad a\psi_x^2 + 2b\psi_x\psi_y + c\psi_y^2 = 0.$$

We obviously must have $a(x_0, y_0) \neq 0$ or $c(x_0, y_0) \neq 0$, say the latter. It suffices to find ψ with $\psi_x(x_0, y_0) \neq 0$ and

$$(I') \quad a + 2b \left(\frac{\psi_y}{\psi_x} \right) + c \left(\frac{\psi_y}{\psi_x} \right)^2 = 0$$

in a neighborhood of (x_0, y_0) ; for we can then take $\phi(x, y) = x$, and $(\phi_x \psi_y - \phi_y \psi_x)(x_0, y_0) = \psi_y(x_0, y_0) \neq 0$. But equation (I') is simply equivalent to

$$\frac{\psi_y}{\psi_x} = -\frac{b}{c} \quad \text{or} \quad c\psi_x = -b\psi_y.$$

This is a first order linear PDE, and the line $y = y_0$ is free near (x_0, y_0) , since $c(x_0, y_0) \neq 0$. So we can find a solution with arbitrary values of $\psi(x, y_0)$ for x near x_0 , in particular with $\psi_x(x_0, y_0) \neq 0$.

If $c(x_0, y_0) = 0$, we look at ψ_x/ψ_y instead. ♦

There is an especially important characterization of elliptic semi-linear PDE's

$$(I) \quad a(x, y)u_{xx} + 2b(x, y)u_{xy} + c(x, y)u_{yy} + \cdots = 0,$$

which is the basis for extending the definition of ellipticity to more general equations. Consider an initial curve c in \mathbb{R}^2 . According to the definitions of section 2, when we are considering the PDE (I), the curve c is free if and only if

$$av_1^2 + 2bv_1v_2 + cv_2^2 \neq 0 \quad \text{on } c$$

(the initial conditions are irrelevant in the semi-linear case). Now if $ac - b^2 > 0$, then the equation

$$0 = a\lambda^2 + 2b\lambda\mu + c\mu^2$$

has no real roots at all, except $\lambda = \mu = 0$. So if (I) is elliptic, then *any* initial curve c is free for any initial conditions. On the other hand, if (I) is hyperbolic, then there is a 2-parameter family of characteristic curves which fail to be free at all points.

For a general second order equation

$$(I) \quad F(x, y, u, p, q, r, s, t) = 0,$$

we define a given solution u to be **elliptic** if

$$4F_r F_t - F_s^2 > 0$$

at all points

$$(x, y, u(x, y), u_x(x, y), u_y(x, y), u_{xx}(x, y), u_{xy}(x, y), u_{yy}(x, y)).$$

In this case, *any* initial curve c is free for the initial data

$$u|_c, \quad u_{x_i}|_c, \quad u_{x_i x_j}|_c.$$

We define a given solution u to be **hyperbolic** or **parabolic** in the obvious analogous way. Notice that for a given PDE which is not semi-linear, and even for a quasi-linear PDE, there may be solutions which are elliptic and also solutions which are hyperbolic or parabolic. The simplest example is the equation

$$u_{xx} + u \cdot u_{yy} = 0.$$

The solution $u = 1$ is elliptic, and the solution $u = -1$ is hyperbolic. This may seem like a ridiculous distinction, but, as we shall learn in sections 8 and 9, solutions near 1 will have entirely different properties from solutions near -1 . A more natural example is the equation

$$(1 - u_x^2)u_{xx} - 2u_x u_y u_{xy} + (1 - u_y^2)u_{yy} = 0,$$

which occurs in gas dynamics. The solution u is elliptic if and only if $u_x^2 + u_y^2 < 1$. Such solutions represent “subsonic” flow, while hyperbolic solutions represent “supersonic” flow. Thus we see that the terms elliptic, hyperbolic, and parabolic, do not make sense for the general second order equation; these terms apply to solutions of the equation, rather than to the equation itself. On the other hand, it clearly makes sense to apply the terms elliptic, hyperbolic, and parabolic to given initial data along a given initial curve.

It is important to observe that the type of a solution remains the same under a change of variable, just as in the semi-linear case. Suppose that u is a solution of

$$(I) \quad F(x, y, u, p, q, r, s, t) = 0,$$

and that we write

$$(V) \quad u(x, y) = v(\phi(x, y), \psi(x, y))$$

for a diffeomorphism (ϕ, ψ) . Using equations (VI), we see that v satisfies an equation

$$(2) \quad G(x, y, v, v_\xi, v_\eta, v_{\xi\xi}, v_{\xi\eta}, v_{\eta\eta}) = 0$$

where G has the form

$$G(—, r, s, t) = F(—, r', s', t')$$

for

$$\begin{aligned} r' &= \phi_x^2 r + 2\phi_x \psi_x s + \psi_x^2 t \\ s' &= \phi_x \phi_y r + (\phi_x \psi_y + \phi_y \psi_x) s + \psi_x \psi_y t \\ t' &= \phi_y^2 r + 2\phi_y \psi_y s + \psi_y^2 t. \end{aligned}$$

Hence

$$G_r = F_r \phi_x^2 + F_s \phi_x \phi_y + F_t \phi_y^2, \quad \text{etc.},$$

and we find that

$$\begin{pmatrix} G_r & \frac{1}{2} G_s \\ \frac{1}{2} G_s & G_t \end{pmatrix} = \begin{pmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{pmatrix} \begin{pmatrix} F_r & \frac{1}{2} F_s \\ \frac{1}{2} F_s & F_t \end{pmatrix} \begin{pmatrix} \phi_x & \psi_x \\ \phi_y & \psi_y \end{pmatrix}.$$

Therefore

$$G_r G_t - \frac{1}{4} G_s^2 = (F_r F_t - \frac{1}{4} F_s^2) \cdot (\phi_x \psi_y - \phi_y \psi_x)^2.$$

Consequently, u is an elliptic or hyperbolic solution of (1) if and only if v is an elliptic or hyperbolic solution of (2).

A few remarks might be made concerning the definitions in higher dimensions (which we do not actually use). A semi-linear PDE

$$\sum_{i,j=1}^n a_{ij}(x_1, \dots, x_n) u_{x_i x_j} + f(x_1, \dots, x_n, u, \dots, u_{x_i}, \dots) = 0$$

is called elliptic at a point $x = (x_1, \dots, x_n)$ if the matrix $(a_{ij}(x))$ is definite, and elliptic in a region if it is elliptic at each point of the region. For an elliptic semi-linear PDE, any initial manifold will be free. We can define several different sorts of hyperbolicity and parabolicity, depending on the rank and signature of this matrix when it is not definite. Proposition 5 generalizes; if the a_{ij} are constants then our equation is equivalent to a normal form

$$u_{x_1 x_1} + \dots + u_{x_r x_r} - u_{x_{r+1} x_{r+1}} - \dots - u_{x_s x_s} + \dots = 0.$$

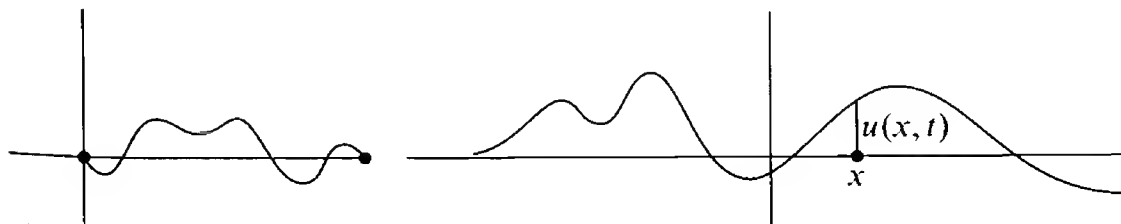
But Theorems 6–8 do not generalize; there are too many conditions to be satisfied by the diffeomorphism which changes u to v . We can also define when a given solution u of the general second order PDE is elliptic, in a fairly obvious way that is left to the reader.

The fact that any initial curve for an elliptic semi-linear second order equation in 2 variables is free, while a hyperbolic semi-linear equation always has initial curves which are characteristic, certainly suggests that elliptic and hyperbolic equations might have quite different properties. But some of the most important reasons for this classification of second order equations come from physics, which provides the motivation for many of the basic problems about them. So we will first make a brief excursion into this forbidding domain.

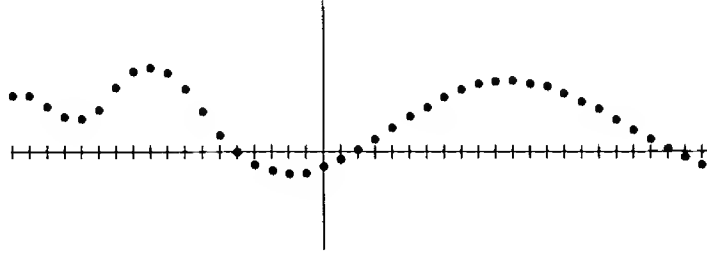
6. THE PROTOTYPICAL PDE's OF PHYSICS

We are going to begin by deriving certain classical PDE's which describe important (somewhat idealized) physical situations. The word “derive” had better be taken with a hefty grain of salt, however. What I have really tried to do is give plausible reasons why the physical situations should be governed by those PDE's which the physicists have agreed upon. I've never really been able to understand which parts of the standard derivations are supposed to be obvious, which are mathematically simplifying assumptions, which steps are supposed to correspond to empirically discovered physical laws, or even what all the words are supposed to mean.

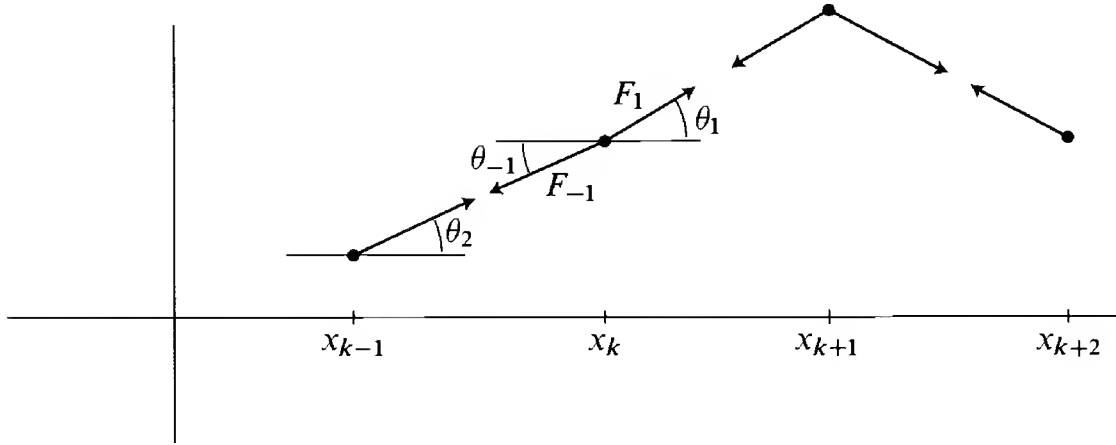
The first idealized physical situation which we want to describe is a vibrating string which is not acted upon by any outside forces. We naturally consider this string to be 1-dimensional, and we will assume that the motion actually takes place in a plane. We may regard our string as being either of finite length with fixed endpoints, or of infinite length. The first possibility corresponds to a string stretched between two prongs, while the second is a more idealized conception. We will let $u(x, t)$ denote the height of the string above $(x, 0)$ at time t .



In order to apply the laws of mechanics to the string, we will first regard it as a discrete collection of point masses, whose x -coordinates are all some small distance h apart. In the course of time, the k^{th} particle moves up and



down (but not sideways); its height above the x -axis at time t will be denoted by $u(x_k, t)$. The only forces acting are the “tension” forces between pairs of particles; physically these come about because the motion of the string involves slight changes in the distances between molecules, to which the intermolecular forces are extremely sensitive. We will assume that each particle is acted upon



only by the particle immediately to its left and right. It is thus influenced by two forces, $F_1(x_k, t)$ and $F_{-1}(x_k, t)$, which are vectors pointing along the line from it to its neighboring particles. These vectors make angles of $\theta_1(x_k, t)$ and $\theta_{-1}(x_k, t)$ with the horizontal rays pointing right and left from the position of the particle. Assuming that our particles all have mass m , we use the law $F = ma$ to obtain the following equation for the vertical motion $u(x_k, t)$ of our particle:

$$(a) \quad m \cdot u_{tt}(x_k, t) = |F_1(x_k, t)| \cdot \sin \theta_1(x_k, t) - |F_{-1}(x_k, t)| \cdot \sin \theta_{-1}(x_k, t).$$

Since the particle does not move sideways, we also have

$$(b) \quad 0 = |F_1(x_k, t)| \cdot \cos \theta_1(x_k, t) - |F_{-1}(x_k, t)| \cdot \cos \theta_{-1}(x_k, t).$$

We clearly have

$$(c) \quad \begin{cases} \cos \theta_1(x_k, t) = \frac{h}{\sqrt{h^2 + \{u(x_{k+1}, t) - u(x_k, t)\}^2}} \\ \quad = \frac{1}{\sqrt{1 + D_k^2}}, \quad \text{where } D_k = \frac{u(x_{k+1}, t) - u(x_k, t)}{h}, \\ \sin \theta_1(x_k, t) = \frac{u(x_{k+1}, t) - u(x_k, t)}{h\sqrt{1 + D_k^2}}. \end{cases}$$

Noting that $\theta_{-1}(x_k, t) = \theta_1(x_{k-1}, t)$, we find that (a)–(c) lead to

$$\begin{aligned} (d) \quad & m \cdot u_{tt}(x_k, t) \\ &= |F_1(x_k, t)| \cdot \left[\sin \theta_1(x_k, t) - \sin \theta_{-1}(x_k, t) \cdot \frac{\cos \theta_1(x_k, t)}{\cos \theta_{-1}(x_k, t)} \right] \\ &= |F_1(x_k, t)| \cdot \left[\frac{u(x_{k+1}, t) - u(x_k, t)}{h\sqrt{1 + D_k^2}} - \frac{u(x_k, t) - u(x_{k-1}, t)}{h\sqrt{1 + D_{k-1}^2}} \cdot \frac{\sqrt{1 + D_{k-1}^2}}{\sqrt{1 + D_k^2}} \right] \\ &= \frac{|F_1(x_k, t)|}{\sqrt{1 + D_k^2}} \cdot \left[\frac{u(x_{k+1}, t) + u(x_{k-1}, t) - 2u(x_k, t)}{h} \right]. \end{aligned}$$

This is a system of differential equations for the (possibly infinitely many) functions $u(x_k, t)$. It depends, of course, on knowing F_1 , which would depend on the particular molecular forces involved. Leaving aside that objection for the moment, we now seek a PDE which will describe a uniform string, not a discrete collection. To obtain this, we want to let the number of particles increase, by decreasing h . Of course, we also want to change m in the process, so as not to have an infinitely heavy string at the end. On a piece of thread of length 1, there will be about $1/h$ particles, with a total mass of m/h . So we keep m/h equal to a constant ρ , the **density**. We will also assume that F_1 approaches a function T , the **tension** of the string (it measures with how much force the string will snap apart if it is cut at some point). Now it is well-known (use Taylor's theorem for a proof) that

$$\lim_{h \rightarrow 0} \frac{f(x+h) + f(x-h) - 2f(x)}{h^2} = f''(x).$$

So if we divide equation (d) by h , and then take the limit as $h \rightarrow 0$, we find that $u(x, t)$ should satisfy

$$\rho u_{tt} = \frac{T}{\sqrt{1 + u_x^2}} u_{xx}.$$

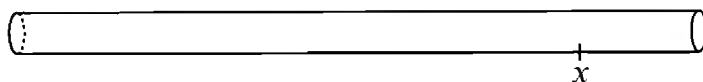
Quite apart from the fact that we don't know how to find T , this equation suffers the defect of being non-linear. We can simplify things by restricting ourselves to the case of *small vibrations*; then T is practically constant, and $\sqrt{1 + u_x^2}$ is practically 1. We have thus completed our devious path to the 1-dimensional "wave equation"

$$u_{xx} = \rho u_{tt}.$$

Since $\rho > 0$, a simple change of coordinates always gives us the equation

$$u_{xx} = u_{tt}.$$

This equation also describes sound waves in a long thin pipe; in this case, $u(x, t)$



represents the density of the air at distance x and time t . The 2-dimensional wave equation

$$u_{xx} + u_{yy} = u_{tt}$$

will describe the motion of a vibrating membrane, while the 3-dimensional wave equation

$$u_{xx} + u_{yy} + u_{zz} = u_{tt}$$

will describe sound waves, as well as certain phenomena involving electromagnetic waves. For our purposes, the 1-dimensional wave equation will be quite adequate.

The second idealized physical situation which we want to describe is the temperature distribution within a material body. It is important here to distinguish between the temperature and the heat energy of a body. The **temperature** of a body, which is operationally defined by putting it in contact with a thermometer, is the average kinetic energy of its molecules. We define the temperature $u(x, y, z)$ of a body B at the point (x, y, z) to be the limit of the temperatures of small parts of B which contain (x, y, z) . Naturally this doesn't really make much sense for a physical body made of molecules, so we must deal with an idealized situation when we consider temperature to be a function u defined on a certain subset $B \subset \mathbb{R}^3$.

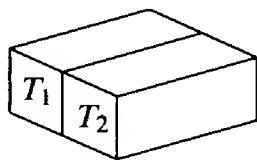
Heat energy is something different. It takes a certain amount of energy to produce a unit increase in temperature within a unit amount of matter. How much energy depends on the particular kind of matter we have. There are two reasons for this. First of all, the molecules in two different kinds of matter have different weights, so different amounts of energy will be required to increase the average kinetic energy of the same number of molecules by the same amount. In perfect gases, this is the only influencing factor. In other cases, the strength of the intermolecular forces will also influence how much energy has to be put in to increase the average kinetic energy. The **specific heat** or **heat capacitance** C of a piece of matter is the amount of energy required to increase the temperature of a unit mass by a unit amount; we will consider only bodies with uniform specific heat. If $u(x, y, z)$ is the temperature at (x, y, z) of an object B with specific heat C and density ρ , then the total **heat energy** of B is

$$(a) \quad \text{heat energy} = C\rho \int_B u \, dV.$$

The basic experimental fact about heat is that when two bodies of different temperature are placed next to each other, the temperature of the hotter one decreases while the temperature of the cooler one increases, and the rate of change of temperature is proportional to the difference. So if we have two bodies which at each time t have uniform temperatures $T_1(t)$ and $T_2(t)$, then

$$\frac{d}{dt} T_1(t) = (\text{constant}) \cdot (T_1(t) - T_2(t)).$$

This constant will depend on the amount of surface area which the two bodies have in common, as well as on the nature of the material of which they are made. The simplest case to consider is that of a single piece of matter B with



two parts, B_1 and B_2 , initially at different temperatures, T_1 and T_2 . Of course, the two parts will not continue to have uniform temperatures, but at least we

can say that

$$\left. \frac{d}{dt} \right|_{t=0} T_1(t) = (\text{constant}) \cdot A \cdot (T_1(t) - T_2(t)),$$

where A is the area between them. By changing the constant, we can also write this as

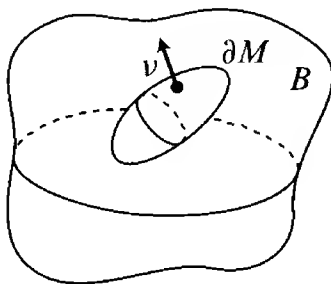
$$(b) \quad \left. \frac{d}{dt} \right|_{t=0} C\rho T_1(t) = \kappa \cdot A \cdot (T_1(t) - T_2(t))$$

for some constant κ . This constant κ is called the **heat conductivity** of the matter in question, since it measures the rate at which heat energy is transferred. Roughly speaking, κ must depend on the way the molecules are arranged; this arrangement will somehow determine to what extent faster moving molecules can influence slower moving ones.

Now let us consider a body $B \subset \mathbb{R}^3$ with uniform density ρ , specific heat C , and heat conductivity κ , but with temperature $u(x, y, z, t)$ varying both with position and time. What will be the analogue of equation (b)? The left side of (b) represents the rate of change of the heat energy of the part of B with temperature T_1 . So for any subset $M \subset B$, equation (a) suggests that the analogue of the left side of (b) is

$$(L) \quad \frac{d}{dt} C\rho \int_M u \, dV = C\rho \int_M u_t \, dV.$$

Let us suppose that $M \subset B$ is a 3-dimensional manifold-with-boundary, and



that ν is the outward unit normal on ∂M . If X is the vector field

$$X = \text{grad } u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z} \right) \quad [X \text{ depends on } t],$$

then for fixed t the function $-\langle X, \nu \rangle$ measures how fast the temperature is decreasing as we cross ∂M from the inside to the outside; roughly speaking it measures the difference $T_1 - T_2$ on the two sides of ∂M . So the analogue of

the right side of (b) is

$$(R) \quad (-\kappa) \int_{\partial M} -\langle X, \nu \rangle dA = \kappa \int_{\partial M} \langle X, \nu \rangle dA.$$

Setting (L) = (R), we are led to the equation

$$(c) \quad C\rho \int_M u_t dV = \kappa \int_{\partial M} \langle X, \nu \rangle dA.$$

We could also obtain equation (c) by breaking M up into small cubes, each of which may be regarded as having constant temperature, and applying (b) to each cube; the term on the right of (b) is to be replaced by a sum over the faces of the cube.

Now applying the Divergence Theorem (Problem I.9-13), we are led from equation (c) to

$$(d) \quad \begin{aligned} C\rho \int_M u_t dV &= \kappa \int_M \operatorname{div} X dV \\ &= \kappa \int_M u_{xx} + u_{yy} + u_{zz} dV. \end{aligned}$$

Since this is supposed to hold for all M , the integrands must be equal, and we obtain the 3-dimensional “heat equation”

$$\frac{\kappa}{C\rho} \cdot (u_{xx} + u_{yy} + u_{zz}) = u_t.$$

Of course, we usually replace the positive constant $\kappa/C\rho$ by 1. The 1-dimensional heat equation

$$u_{xx} = u_t$$

describes temperature distribution in a long thin rod, while the 2-dimensional heat equation

$$u_{xx} + u_{yy} = u_t$$

describes temperature distribution in a thin plate.

We obtain a very special equation when we seek the “steady state” temperature distribution of a body. This is the temperature distribution it has when the temperature is *not* varying with time. For example, if we keep both ends of a bar at fixed temperatures by attaching them to “heat reservoirs”, mechanisms which maintain a fixed temperature at a point, then the temperature distribution will rapidly approach a linear function between these two values. To find the steady state temperature distribution, we just set $u_t = 0$ in the heat equation. Thus in

the 1-dimensional case we obtain simply $u_{xx} = 0$, whose solutions are simply linear functions on \mathbb{R} . In the 2- and 3-dimensional cases, we obtain

$$\begin{aligned} u_{xx} + u_{yy} &= 0 && \text{2-dimensional Laplace equation} \\ u_{xx} + u_{yy} + u_{zz} &= 0 && \text{3-dimensional Laplace equation.} \end{aligned}$$

Among these important equations of mathematical physics, we find representatives of each of the three types of second order PDE's. In particular, for two variables we have the following standard examples:

Elliptic equations

$u_{xx} + u_{yy} = 0$	the 2-dimensional Laplace equation
-----------------------	------------------------------------

Hyperbolic equations

$u_{xx} - u_{yy} = 0$	the 1-dimensional wave equation
-----------------------	---------------------------------

The equation $u_{xx}(x, y) = 0$ is a parabolic equation in 2 variables, but obviously a little too simple to be very representative. The standard representative is

Parabolic equations

$u_{xx} = u_y$	the 1-dimensional heat equation
----------------	---------------------------------

Parabolic equations are often slighted in introductory treatments of PDE's, and they will suffer the same treatment in our hands—we will never look at them again. We therefore say good-bye to the heat equation, and consider only the special case of Laplace's equation.

Now let us see what sort of mathematical questions these physical situations suggest. Consider first the 1-dimensional wave equation, which we will write as $u_{xx} - u_{tt} = 0$, to remind us that it describes a process involving time. In such processes, it is naturally of interest to predict what will happen later from a knowledge of what is happening now. It seems perfectly reasonable to hope that we can predict the motion of a string in terms of its initial position and initial velocity,

$$u(x, 0), \quad u_t(x, 0).$$

Moreover, there seems to be no reason why we should have to limit ourselves to analytic initial conditions. For example, if we “pluck” a string, then the simplest description involves an initial condition $u(x, 0)$ which is not even differentiable



everywhere. This example suggests that for *hyperbolic* equations the Cauchy problem is the right one to pose, and that we should not have to restrict ourselves to analytic initial data (which is all we can treat when we rely on the Cauchy-Kowalewski Theorem).

Another mathematical problem is suggested by the fact that in actuality vibrating strings are always secured at two ends. Given two functions $\phi, \psi: [0, L] \rightarrow \mathbb{R}$ with

$$\begin{aligned}\phi(0) &= \phi(L) = 0 \\ \psi(0) &= \psi(L) = 0,\end{aligned}$$

we can ask for a solution u of the 1-dimensional wave equation with

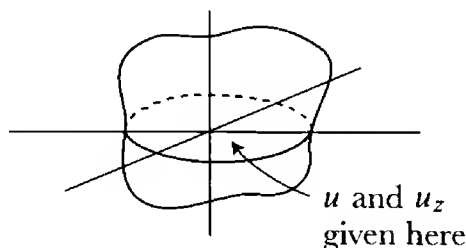
$$\begin{aligned}u(x, 0) &= \phi(x) & 0 \leq x \leq L \\ u_t(x, 0) &= \psi(x) & 0 \leq x \leq L \\ u(0, t) &= u(L, t) = 0 & \text{for all } t.\end{aligned}$$

This is an example of an “initial-boundary value” problem; although such problems are also quite important, we will not consider them at all.

Quite different questions are suggested by Laplace’s equation

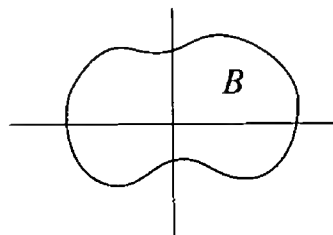
$$u_{xx} + u_{yy} + u_{zz} = 0.$$

Here time is not involved at all; the equation describes the steady state heat distribution of some object $B \subset \mathbb{R}^3$. The Cauchy problem for this equation would correspond to the physical problem of predicting the temperature everywhere in B from a knowledge of its values along some plane in B , together with knowledge of its derivative in the perpendicular direction. Now this is hardly a reasonable problem, since it isn’t very easy to measure the temperature at various points inside a solid object. This is the sort of information we would



like to *predict*. The sort of thing we *can* measure is the temperature along the boundary. Similarly, for a region B in \mathbb{R}^2 , we would like to find the solutions u of the 2-dimensional Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$



which has given values along the boundary of B . A problem of this sort is called a **Dirichlet problem**. The physics seems to suggest that for elliptic equations it is the Dirichlet problem rather than the Cauchy problem which should be of interest.

Now let us see how these physical speculations correspond to mathematical reality.

The 1-dimensional wave equation

$$(1) \quad u_{xx} - u_{yy} = 0$$

is admirably suited to illustrate the general behavior of hyperbolic equations, because the most general solution of (1) can be written down completely. The trick for doing this is simply to use the alternative standard form for a hyperbolic equation. We define v by

$$(2) \quad v(\xi, \eta) = u\left(\frac{\xi + \eta}{2}, \frac{\xi - \eta}{2}\right) \quad u(x, y) = v(x + y, x - y).$$

Then equation (1) for u gives

$$(3) \quad v_{\xi\eta} = 0.$$

At the very beginning of this chapter we mentioned that the general solution of equation (3) is

$$v(\xi, \eta) = f(\xi) + g(\eta).$$

So the general solution of (1) is

$$(4) \quad \begin{aligned} u(x, y) &= v(x + y, x - y) \\ &= f(x + y) + g(x - y). \end{aligned}$$

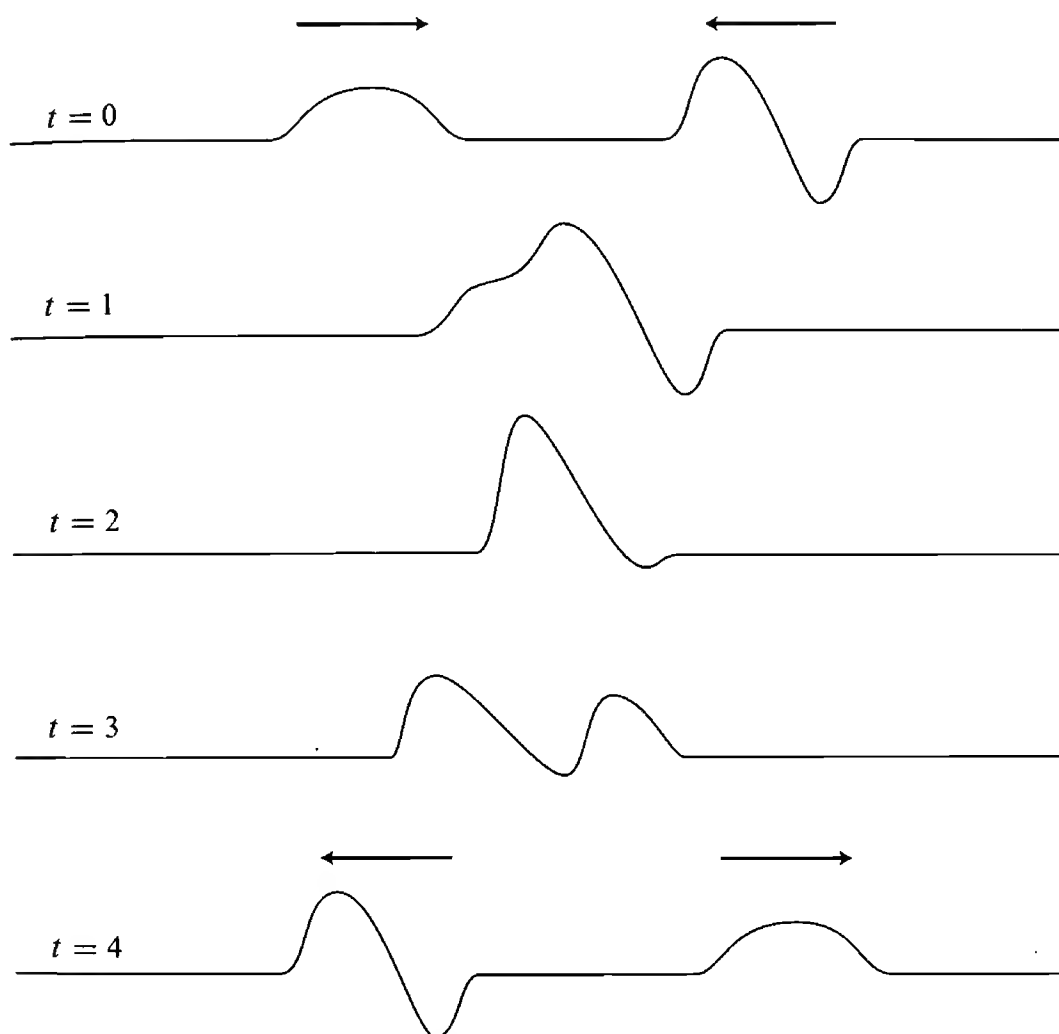
If we think of our equation in terms of position x and time t ,

$$u_{xx} - u_{tt} = 0,$$

then the solution

$$u(x, t) = f(x + t) + g(x - t)$$

represents the sum of two “waves”, the first moving to the left as t increases, the second moving to the right.



Using the representation (4) for the solution u of (1), it is easy to find solutions with given initial conditions

$$(5) \quad \begin{cases} u(x, 0) = \phi(x) \\ u_y(x, 0) = \psi(x). \end{cases}$$

Clearly we must have

$$(6) \quad \begin{cases} f(x) + g(x) = \phi(x) \\ f'(x) - g'(x) = \psi(x), \end{cases}$$

and therefore

$$\begin{cases} f'(x) + g'(x) = \phi'(x) \\ f'(x) - g'(x) = \psi(x), \end{cases}$$

which implies

$$f'(x) = \frac{\phi'(x) + \psi(x)}{2}, \quad g'(x) = \frac{\phi'(x) - \psi(x)}{2}.$$

This means that we must have

$$(7) \quad f(x) = \frac{\phi(x)}{2} + \frac{1}{2} \int_0^x \psi(s) ds + C_1, \quad g(x) = \frac{\phi(x)}{2} - \frac{1}{2} \int_0^x \psi(s) ds + C_2$$

for certain constants C_1, C_2 ; and to satisfy (6) we must have $C_1 = -C_2$. Using (4) we then find that u must be

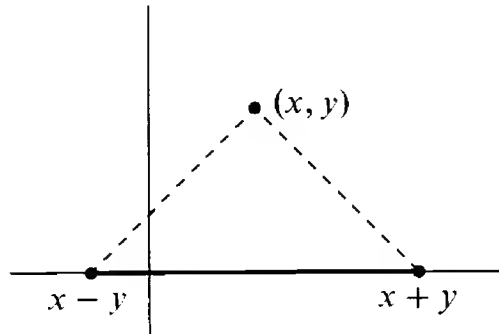
$$(8) \quad u(x, y) = \frac{\phi(x+y) + \phi(x-y)}{2} + \frac{1}{2} \int_{x-y}^{x+y} \psi(s) ds.$$

It is clear, moreover, that this u is a solution of equation (1), with initial conditions (5).

Notice that the boundary values ϕ, ψ enter into the solution in quite a different way than for first order PDE's. If u is a solution of

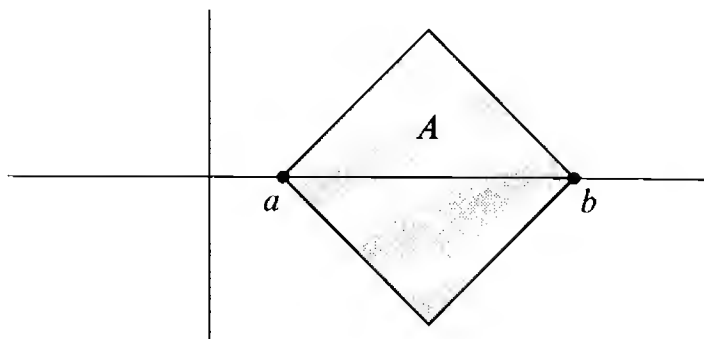
$$F(x, y, u, p, q) = 0$$

with initial data $\bar{u}, \bar{p}, \bar{q}$ along a free curve σ , then the value $u(x, y)$ of u at a particular point (x, y) depends on the value of the initial data at one point (\bar{x}, \bar{y}) on σ , namely the intersection of σ with the base curve of the characteristic strip through $(x, y, u(x, y), u_x(x, y), u_y(x, y))$. Changing the initial data on an interval which does not contain (\bar{x}, \bar{y}) will not change the value of the solution u at (x, y) . But in the solution (8) of the 1-dimensional wave equation (1), we need to know the values of ϕ and ψ on the whole interval $[x-y, x+y]$ (or $[x+y, x-y]$ if $y < 0$). This interval is therefore called the “domain of dependence” of the point (x, y) .



dence” of the point (x, y) . Conversely, if we are given initial conditions ϕ, ψ

defined only on an interval $[a, b]$, then equation (8) defines u only on the set A of all points (x, y) whose domain of dependence is contained in $[a, b]$. Notice



that A is bounded by the curves through a and b which are characteristic for the PDE (1).

Naturally, we might ask about solutions of (1) along free curves other than the x -axis. In section 8 we will consider this question for even more general hyperbolic equations. If the 1-dimensional wave equation is indeed representative of general hyperbolic equations, then we should be able to solve the Cauchy problem for any hyperbolic equation, along any free curve, and without any assumptions about analyticity of the initial conditions.

The situation is completely different for the 2-dimensional Laplace equation

$$(9) \quad u_{xx} + u_{yy} = 0.$$

Solutions of this equation are called **harmonic** functions (on \mathbb{R}^2), and, as we have pointed out in Chapter 9, their study is closely related to the theory of complex analytic functions. In a simply-connected open subset of \mathbb{R}^2 , every harmonic function u is the real part of a complex analytic function $u + iv$; and conversely, the real part of a complex analytic function is always harmonic. This means, in particular, that *every* solution of (9) is automatically real analytic on \mathbb{R}^2 . So we cannot hope to solve equation (9) with initial conditions

$$(10) \quad \begin{cases} u(x, 0) = \phi(x) \\ u_y(x, 0) = \psi(x) \end{cases}$$

in a neighborhood of the x -axis unless ϕ and ψ are real analytic.* Moreover, if ϕ and ψ are real analytic, then the problem of finding a solution of (9) with initial conditions (10) is essentially trivial. We note that if $u + iv$ is analytic, then

*However, there may be solutions in the upper or lower half-plane.

the Cauchy-Riemann equations give $v_x = -u_y$. So the initial conditions allow us to determine v_x along the x -axis, and therefore determine v up to a constant along the x -axis. So the complex analytic function $u + iv$ is determined up to an imaginary constant on the x -axis, which means that $u + iv$ is determined up to an imaginary constant on the plane.

These remarks really amount to a restatement of the fact, already observed in the proof of the Cauchy-Kowalewski Theorem, that for analytic equations with analytic data, the coefficients of the presumptive analytic solution are easily determined. It is perhaps of interest to note that we can formally solve (9) by analogy with the wave equation (1). If we formally define

$$(11) \quad v(\xi, \eta) = u\left(\frac{\xi + \eta}{2}, \frac{\xi - \eta}{2i}\right) \quad u(x, y) = v(x + iy, x - iy),$$

then equation (9) becomes $v_{\xi\eta} = 0$, which leads us to

$$\begin{aligned} v(\xi, \eta) &= f(\xi) + g(\eta) \\ u(x, y) &= f(x + iy) + g(x - iy). \end{aligned}$$

Taking into account the initial conditions (10), we are led to the formal solution

$$(12) \quad u(x, y) = \frac{\phi(x + iy) + \phi(x - iy)}{2} + \frac{1}{2} \int_{x-iy}^{x+iy} \psi(z) dz.$$

If ϕ and ψ are real analytic, and hence have complex analytic extensions, then this formula makes sense—the integral may be taken along any path from $x - iy$ to $x + iy$. Because ϕ and ψ are real on the real axis, the function u is real-valued, and is easily seen to satisfy (9) and (10).

Our physical considerations suggest that we should be able to solve the Dirichlet problem for (9): given a function $f: \partial B \rightarrow \mathbb{R}$ on the boundary of a region $B \subset \mathbb{R}^2$, we ought to be able to find a solution u of (9) with $u = f$ on ∂B . In complex analysis courses it is shown that this is indeed the case.

7. HYPERBOLIC SYSTEMS IN TWO VARIABLES

In this section we will consider first order quasi-linear systems in two variables. Our initial manifolds for the Cauchy problem will therefore be curves in \mathbb{R}^2 , and we are naturally only interested in initial data for which the initial curve is free. So without loss of generality, we assume that our initial curve is an interval $[a, b]$ of the x -axis, and that our quasi-linear system of n equations

for n unknown functions $u^i: \mathbb{R}^2 \rightarrow \mathbb{R}$ is of the form

$$u_y^i(x, y) = \sum_{j=1}^n a_{ij}(x, y, u^1(x, y), \dots, u^n(x, y)) \cdot u_x^j(x, y) + b^i(x, y, u_1(x, y), \dots, u_n(x, y)).$$

We will often consider $u = (u^1, \dots, u^n)$ to be a column vector, just so that we can multiply on the left by a matrix. Then we can write our system as

$$u_y(x, y) = A(x, y, u(x, y)) \cdot u_x(x, y) + b(x, y, u(x, y))$$

where A is an $n \times n$ matrix of functions, and b is a column vector. More briefly, we have the system

$$u_y = A \cdot u_x + b.$$

This system is called **hyperbolic for given initial conditions** $\overset{\circ}{u} = (\overset{\circ}{u}^1, \dots, \overset{\circ}{u}^n)$ on an interval $[a, b]$ of the x -axis if A is diagonalizable in a neighborhood of all points $(x, 0, \overset{\circ}{u}(x))$ for $x \in [a, b]$; more precisely, it is C^k **hyperbolic** if there is a C^k matrix T such that TAT^{-1} is diagonalizable [this more precise formulation is necessary, because even if A is C^k and always diagonalizable, it may not be possible to choose the diagonalizing matrix T to be C^k]. The basic result is that a quasi-linear system with hyperbolic initial conditions has a solution with these initial conditions; in the next section we will apply this to a single second order equation.

In order to explain the main points of the argument, we will first sketch how the proof would go in the “semi-linear” case,

$$(a) \quad u_y(x, y) = A(x, y)u_x(x, y) + b(x, y, u(x, y)),$$

where A does not depend on u (although b might). Let T be the matrix which diagonalizes A , and define $v = (v^1, \dots, v^n)$ by

$$u(x, y) = T(x, y) \cdot v(x, y),$$

so that

$$u_x = T_x v + T v_x, \quad u_y = T_y v + T v_y.$$

Substituting into (a), we obtain

$$T_y v + T v_y = AT_x v + AT v_x + b,$$

so

$$(b) \quad \begin{aligned} v_y &= (T^{-1}AT)v_x + (T^{-1}AT_x v + T^{-1}b - T^{-1}T_y v) \\ &= Cv_x + d. \end{aligned}$$

Clearly u satisfies (a) with initial conditions $\overset{\circ}{u}$ if and only if v satisfies (b) with initial conditions

$$\overset{\circ}{v}(x) = T^{-1}(x, 0) \cdot \overset{\circ}{u}(x).$$

So we might as well assume that we have the equation

$$(l) \quad u_y = Cu_x + d,$$

where the matrix

$$C(x, y) = \begin{pmatrix} \lambda^1(x, y) & & 0 \\ & \ddots & \\ 0 & & \lambda^n(x, y) \end{pmatrix}$$

is a diagonal matrix. Thus our equation reads

$$(l') \quad -\lambda^i(x, y) \cdot u_x^i + u_y^i = d^i(x, y, u^1(x, y), \dots, u^n(x, y)).$$

The vector field

$$X_i(x, y) = (-\lambda^i(x, y), 1)$$

is called the i^{th} **characteristic vector field** of equation (l'), and the integral curves of X_i are the i^{th} family of **characteristic curves**. We might as well consider only characteristic curves of the form $t \mapsto (c(t), t)$. If u is a solution of (l') and $t \mapsto (c(t), t)$ is a characteristic curve of the i^{th} family, then

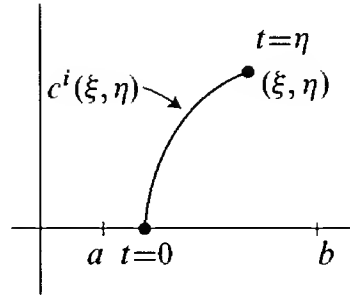
$$(2) \quad \begin{aligned} \frac{d}{dt} u^i(c(t), t) &= -\lambda^i(c(t), t) \cdot u_x^i(c(t), t) + u_y^i(c(t), t) \\ &= d^i(c(t), t, u(c(t), t)). \end{aligned}$$

Consequently,

$$(3) \quad u^i(c(\eta), \eta) = u^i(c(0), 0) + \int_0^\eta d^i(c(t), t, u(c(t), t)) dt.$$

In particular, let $t \mapsto (c^i(\xi, \eta), t)$ be the characteristic curve of the i^{th} family which satisfies

$$(4) \quad c^i(\xi, \eta)(\eta) = \xi.$$



If $c^i(\xi, \eta)(0) \in [a, b]$, then equations (3) and (4) give

$$(5) \quad u^i(\xi, \eta) = \bar{u}^i(c^i(\xi, \eta)(0)) + \int_0^\eta d^i(c^i(\xi, \eta)(t), t, u(c^i(\xi, \eta)(t), t)) dt.$$

Conversely, if u satisfies (5) in a region obtained by following all characteristic curves from $[a, b] \times \{0\}$ for a certain time interval in either direction, then u will be a solution of (1') with initial conditions \bar{u} on $[a, b] \times \{0\}$. Equation (5) is something like the integral equation which we solved in Chapter I.5, when we proved that differential equations have solutions, but it is more complicated, because the different components of u are integrated over different curves. Nevertheless, it can be solved in essentially the same way. We define an operator S which takes u to the n -tuple of functions Su given by

$$(Su)^i(\xi, \eta) = \bar{u}^i(c^i(\xi, \eta)(0)) + \int_0^\eta d^i(c^i(\xi, \eta)(t), t, u(c^i(\xi, \eta)(t), t)) dt.$$

Then we show that on a suitable complete space of functions the operator S is a contraction, so that it has a fixed point.

When we look at the quasi-linear system

$$(1) \quad u_y(x, y) = A(x, y, u(x, y)) \cdot u_x(x, y) + b(x, y, u(x, y)),$$

we run into a problem at the very first step. For the matrix T which diagonalizes A will depend on u . If we set

$$\begin{aligned} [T]_x &= \frac{\partial T(x, y, u(x, y))}{\partial x} = T_x + \sum_j T_{u^j} u_x^j \\ [T]_y &= \frac{\partial T(x, y, u(x, y))}{\partial y} = T_y + \sum_j T_{u^j} u_y^j, \end{aligned}$$

then the substitution $u(x, y) = T(x, y, u(x, y)) \cdot v(x, y)$ leads to equation (b) again, except that now T_x and T_y are replaced by $[T]_x$ and $[T]_y$, which involve u ; so we do not even obtain an equation for v . We can reduce our system

to one in diagonal form by means of a somewhat more complicated substitution; however, it will be necessary to assume that the matrix $A(x, y, u(x, y))$ in (1) is invertible [in a neighborhood of the points $(x, 0, \overset{\circ}{u}(x))$].

Suppose we have u satisfying (1), and $T(x, y, u)$ diagonalizes $A(x, y, u)$ for all (x, y, u) in a neighborhood of the points $(x, 0, \overset{\circ}{u}(x))$. Define v by

$$(2) \quad u_y = Tv \quad (\text{i.e., } u_y(x, y) = T(x, y, u(x, y)) \cdot v(x, y)).$$

Then

$$(3) \quad Tv = Au_x + b$$

$$\Downarrow$$

$$(4) \quad u_x = A^{-1}(Tv - b).$$

Differentiating (3) with respect to y , we obtain

$$\begin{aligned} Tv_y + [T]_y v &= Au_{xy} + [A]_y u_x + [b]_y \\ &= A(Tv)_x + [A]_y u_x + [b]_y \quad \text{by (2)} \\ &= A[T]_x v + ATv_x + [A]_y u_x + [b]_y, \end{aligned}$$

so

$$v_y = (T^{-1}AT)v_x + T^{-1}A[T]_x v + T^{-1}[A]_y u_x + T^{-1}[b]_y - T^{-1}[T]_y v.$$

Writing out $[T]_x$, $[A]_y$, \dots , substituting for the u_y from (2) and for the u_x from (4), we obtain

$$\begin{aligned} (5) \quad v_y &= (T^{-1}AT)v_x + T^{-1}A \left[T_x + \sum_j T_{u^j} \{A^{-1}(Tv - b)\}^j \right] v \\ &\quad + T^{-1} \left[A_y + \sum_j A_{u^j} (Tv)^j \right] A^{-1}(Tv - b) \\ &\quad + T^{-1} \left[b_y + \sum_j b_{u^j} (Tv)^j \right] - T^{-1} \left[T_y + \sum_j T_{u^j} (Tv)^j \right] v. \end{aligned}$$

If w is the column vector $u^1, \dots, u^n, v^1, \dots, v^n$, then equations (2) and (5) together can be written in the form

$$w_y = Cw_x + d$$

where C is *diagonal*. We have the initial conditions

$$(6) \quad w(x, 0) = \begin{pmatrix} \overset{\circ}{u}(x) \\ T^{-1}(x, 0, \overset{\circ}{u}(x)) \cdot [A(x, 0, \overset{\circ}{u}(x)) \cdot \overset{\circ}{u}'(x) + b(x, 0, \overset{\circ}{u}(x))] \end{pmatrix}.$$

Note that if A , b and u are C^2 , and A is C^2 diagonalizable, then C and d are C^1 ; if the initial condition $\overset{\circ}{u}$ is C^2 , then the initial condition for w is C^1 . Conversely, we have

9. LEMMA. Let $w = (u^1, \dots, u^n, v^1, \dots, v^n)$ be a C^1 solution of the system (2), (5) with C^1 coefficients C and d and C^1 initial conditions (6). Then u satisfies $u_y = Au_x + b$ (and is C^2).

PROOF. Substituting (2) into the last three terms of (5), multiplying by T , and rearranging, we find that

$$(7) \quad Tv_y + [T]_y v - [b]_y \\ = ATv_x + A \left[T_x + \sum_j T_{u^j} \{A^{-1}(Tv - b)\}^j \right] v + [A]_y A^{-1}(Tv - b).$$

Equation (2) implies that u_y has a continuous partial derivative with respect to x . Hence, by a theorem of calculus, u_x has a continuous first partial derivative with respect to y and $u_{xy} = u_{yx}$. Thus

$$(8) \quad u_{xy} = u_{yx} = (Tv)_x = Tv_x + [T]_x v.$$

Define

$$s = A^{-1}(Tv - b) - u_x.$$

Then s has a continuous first partial derivative with respect to y , and

$$\begin{aligned} s_y &= [A^{-1}]_y(Tv - b) + A^{-1}(Tv_y + [T]_y v - [b]_y) - u_{xy} \\ &= -A^{-1}[A]_y A^{-1}(Tv - b) + A^{-1}(Tv_y + [T]_y v - [b]_y) - u_{xy} \\ &= -A^{-1}[A]_y(s + u_x) + A^{-1}(Tv_y + [T]_y v - [b]_y) - u_{xy}. \end{aligned}$$

Multiplying by A we have

$$\begin{aligned} As_y &= -[A]_y s - [A]_y u_x + (Tv_y + [T]_y v - [b]_y) - Au_{xy} \\ &= -[A]_y s - [A]_y u_x \\ &\quad + ATv_x + A \left[T_x + \sum_j T_{u^j} \{A^{-1}(Tv - b)\}^j \right] v + [A]_y A^{-1}(Tv - b) \\ &\quad - A(Tv_x + [T]_x v) \quad \text{by (7) and (8)} \\ &= -[A]_y s - [A]_y u_x + A \left[T_x + \sum_j T_{u^j} s^j + \sum_j T_{u^j} u_x^j \right] v \\ &\quad + [A]_y(s + u_x) - A[T]_x v \\ &= A \left(\sum_j T_{u^j} s^j \right) v. \end{aligned}$$

Thus

$$s_y = \sum_j (T_{u^j} s^j) v.$$

For fixed x , this is a system of ordinary differential equations. But the initial conditions (6) show that $s^j(x, 0) = 0$. So by uniqueness of solutions, we have $s = 0$, which means that

$$\begin{aligned} Au_x + b &= Tv \\ &= u_y \quad \text{by (2).} \end{aligned}$$

Since v is C^1 , the partial derivative $u_y = Tv$ is C^1 ; hence $u_x = A^{-1}(u_y - b)$ is also C^1 . Thus u is C^2 . ♦

Because of Lemma 9, we now restrict our attention to equations

$$u_y(x, y) = C(x, y, u(x, y)) \cdot u_x(x, y) + d(x, y, u(x, y))$$

where $C(x, y, z)$ is a diagonal $n \times n$ matrix in a neighborhood of the points $(x, 0, \overset{\circ}{u}(x))$. One further simplification is possible. Introduce two new unknowns u^{n+1}, u^{n+2} , and consider the equations

$$\begin{cases} u_y(x, y) = C(u^{n+1}(x, y), u^{n+2}(x, y), u(x, y)) \cdot u_x(x, y) \\ \quad + d(u^{n+1}(x, y), u^{n+2}(x, y), u(x, y)) \\ u_y^{n+1}(x, y) = 0 \\ u_y^{n+2}(x, y) = 1 \end{cases}$$

with the initial conditions

$$\begin{cases} u(x, 0) = \overset{\circ}{u}(x) \\ u^{n+1}(x, 0) = x \\ u^{n+2}(x, 0) = 0. \end{cases}$$

A solution u, u^{n+1}, u^{n+2} of this system clearly gives a solution u of the original equation, with the initial conditions $u(x, 0) = \overset{\circ}{u}$. So we might as well consider an equation of the form

$$(1) \quad u_y(x, y) = C(u(x, y)) \cdot u_x(x, y) + d(u(x, y)),$$

with initial conditions

$$(1_0) \quad u(x, 0) = \overset{\circ}{u}(x),$$

where the matrix $C = \begin{pmatrix} \lambda^1 & & 0 \\ & \ddots & \\ 0 & & \lambda^n \end{pmatrix}$ is a diagonal matrix in a neighborhood of the points $\overset{\circ}{u}(x)$.

Our procedure for solving equation (1) is somewhat more involved than the procedure outlined in the semi-linear case since C now depends on u . For any function u , we define the i^{th} family of **characteristic curves** of u to be the curves $t \mapsto (c(t), t)$ with

$$\frac{dc(t)}{dt} = -\lambda^i(u(c(t), t)).$$

Let $c^i(u; \xi, \eta)$ satisfy this equation and the initial condition

$$c^i(u; \xi, \eta)(\eta) = \xi.$$

As before, we find that if u is a solution of (1), with initial conditions (l_0) , then

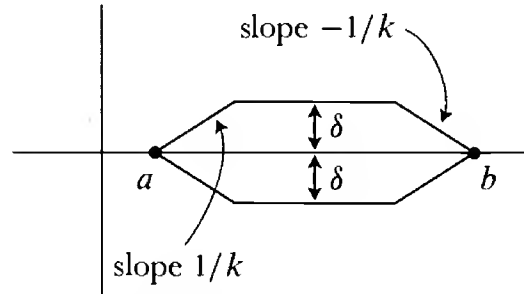
$$(2) \quad u^i(\xi, \eta) = \overset{\circ}{u}^i(c^i(u; \xi, \eta)(0)) + \int_0^\eta d^i(u(c^i(u; \xi, \eta)(t), t)) dt.$$

And, conversely, if u satisfies (2), then u will satisfy (1), with initial conditions (l_0) . We are thus led to define an operator S which takes u to the n -tuple of functions Su given by

$$(Su)^i(\xi, \eta) = \overset{\circ}{u}^i(c^i(u; \xi, \eta)(0)) + \int_0^\eta d^i(u(c^i(u; \xi, \eta)(t), t)) dt.$$

The problem is to show that on a suitable complete metric space of functions, the operator S is a contraction; its fixed point will then be a solution of our equation. The proof is carried out in detail in Courant and Lax [1]. It involves a series of estimates that only an analyst could love, and there doesn't seem to be much point reproducing it here, since the paper is readily accessible, and the sane differential geometer would probably skip it anyway. We would simply like to give a precise statement. Let K be a constant such that all $|\lambda^i| \leq K$ on a region containing $[a, b] \times \{0\}$. We consider the region $\Delta(\delta)$ in \mathbb{R}^2 bounded by the lines

$$\begin{aligned} y &= \delta, & y &= -\delta \\ y &= \frac{1}{K}(x - a), & y &= -\frac{1}{K}(x - a) \\ y &= -\frac{1}{K}(x - b), & y &= \frac{1}{K}(x - b) \end{aligned}$$



(then the characteristic curves of a function u will have slopes which are larger than the slopes of the sides of $\Delta(\delta)$; so if a characteristic curve begins in $\Delta(\delta)$, it will stay in $\Delta(\delta)$ until it hits the x -axis).

10. THEOREM. Consider a system, with initial conditions,

$$\begin{aligned} u_y &= Au_x + b \\ u_x(x, 0) &= \overset{\circ}{u}(x) \quad x \in [a, b], \end{aligned}$$

such that either

- (1) the system is semi-linear, A , b , and $\overset{\circ}{u}$ have continuous partial derivatives satisfying a Lipschitz condition, and A is diagonalizable by a matrix T with the same property

or

- (2) the system is quasi-linear, A is invertible, A , b , and $\overset{\circ}{u}$ have continuous second partial derivatives satisfying a Lipschitz condition, and A is diagonalizable by a matrix T with the same property.

Then for sufficiently small $\delta > 0$ (which depends on the constants in the Lipschitz conditions), there is a unique solution of the system in the region $\Delta(\delta)$. In case (1), the solution u has continuous partial derivatives satisfying a Lipschitz condition, and in case (2), the solution u has continuous second partial derivatives satisfying a Lipschitz condition.

We are stating this particular theorem simply because it is the most accessible in the literature. Other approaches allow all sorts of improvements. First of all, the differentiability requirements can be weakened. Second of all, the matrix A need not be invertible even in the quasi-linear case—but then the approach has to be changed considerably. More important, we would like to know that the solutions have a high degree of differentiability if the coefficients and initial conditions do. The proof of this requires considerations like those which are used to prove that the solutions of an ordinary differential equation are differentiable in the initial conditions, considerations which we already omitted in Volume I. We would also like to consider systems depending on parameters, and show that the solutions are differentiable in the parameters; in section 9 we will use this fact. As in the case of ordinary differential equations, differentiability in the parameters is not very hard, and readers may work this out for themselves, guided by Problem I.5-5.

Despite the somewhat unsatisfactory state in which this section ends, I hope the reader will feel fairly convinced that hyperbolic systems have solutions. In the next two sections we will use this fact to show the enormous difference between hyperbolic and elliptic solutions of second order equations.

8. HYPERBOLIC SECOND ORDER EQUATIONS IN TWO VARIABLES

Consider a second order equation

$$(I) \quad 0 = F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = F(x, y, u, p, q, r, s, t),$$

an initial curve, and hyperbolic initial data along this curve such that the curve is free for the initial data. As we saw in section 2, by introducing a diffeomorphism of the plane, we can assume that the initial curve is a segment $[a, b]$ of the x -axis. The initial data then amount to functions

$$\overset{\circ}{u}, \overset{\circ}{q}, \overset{\circ}{t}$$

on $[a, b]$ satisfying

$$(I_0) \quad 0 = F(x, 0, \overset{\circ}{u}(x), \overset{\circ}{u}'(x), \overset{\circ}{q}(x), \overset{\circ}{u}''(x), \overset{\circ}{q}'(x), \overset{\circ}{t}(x)),$$

and the initial data will still be hyperbolic,

$$F_s^2 - 4F_r F_t > 0,$$

as we remarked in section 5. Setting

$$\alpha(x) = (x, 0, \overset{\circ}{u}(x), \overset{\circ}{u}'(x), \overset{\circ}{q}(x), \overset{\circ}{u}''(x), \overset{\circ}{q}'(x), \overset{\circ}{t}(x)),$$

the requirement that the initial curve $[a, b] \times \{0\}$ be free means that

$$0 \neq F_t(\alpha(x)), \quad \text{for } x \in [a, b].$$

We claim that we can also assume that $0 \neq F_r(\alpha(x))$. The reason for this is that $F_r = 0$ precisely when the direction of the y -axis is characteristic, and we can always avoid this by an appropriate transformation of the plane. In detail (compare page 53), define a new function v by

$$u(x, y) = v(x + \lambda y, y),$$

where λ is a constant. Note that the map $(x, y) \mapsto (x + \lambda y, y)$ takes the segment $[a, b]$ of the x -axis into itself. Now

$$\begin{array}{ll} u_x = v_x, & u_y = \lambda v_x + v_y \\ u_{xx} = v_{xx} & \\ u_{xy} = \lambda v_{xx} + v_{xy} & \\ u_{yy} = \lambda^2 v_{xx} + 2\lambda v_{xy} + v_{yy} & \end{array} \quad \begin{array}{l} \text{[all partials of } u \text{ evaluated} \\ \text{at } (x, y), \\ \text{all partials of } v \text{ evaluated} \\ \text{at } (x + \lambda y, y)]. \end{array}$$

So equation (I) is equivalent to the equation

$$0 = F(x - \lambda y, y, v, v_x, \lambda v_x + v_y, v_{xx}, \lambda v_{xx} + v_{xy}, \lambda^2 v_{xx} + 2\lambda v_{xy} + v_{yy})$$

[all functions v, \dots, v_{yy} evaluated at (x, y)].

This can be written

$$0 = G(x, y, v, v_x, v_y, v_{xx}, v_{xy}, v_{yy}),$$

where G has the form

$$G(_, r, s, t) = F(_, r, \lambda r + s, \lambda^2 r + 2\lambda s + t).$$

Then we have (leaving out the arguments for convenience)

$$G_r = F_r + \lambda F_s + \lambda^2 F_t.$$

By choosing λ sufficiently large we can insure that $G_r(\alpha(x)) \neq 0$ for all $x \in [a, b]$. So we can assume that $F_r(\alpha(x)), F_t(\alpha(x)) \neq 0$ for $x \in [a, b]$.

One way of treating equation (I) would be to first reduce it to an equivalent one by the considerations of section 2. Since $F_t \neq 0$, there is a function f , defined in a neighborhood of all points

$$\beta(x) = (x, 0, \overset{\circ}{u}(x), \overset{\circ}{u}'(x), \overset{\circ}{q}(x), \overset{\circ}{u}''(x), \overset{\circ}{q}'(x)),$$

such that

$$(a) \quad \overset{\circ}{t}(x) = f(\beta(x))$$

$$(b) \quad F(x, y, u, p, q, r, s, f(x, y, u, p, q, r, s)) = 0.$$

So equation (I) is equivalent to the equation

$$(I') \quad u_{yy} = f(x, y, u, u_x, u_y, u_{xx}, u_{xy}).$$

Differentiating (b) with respect to r and s we obtain

$$0 = F_r + F_t f_r$$

$$0 = F_s + F_t f_s,$$

so we have

$$F_r \neq 0 \implies f_r \neq 0$$

$$F_s^2 - 4F_r F_t > 0 \implies f_s^2 + 4f_r > 0$$

(the latter is nothing more than the condition that the solution u of the equation $f(x, y, u, p, q, r, s) - t = 0$ be hyperbolic).

In section 3 we showed that the Cauchy problem for equation (I') is equivalent to a Cauchy problem for the system

$$\begin{aligned}
 & u_y = v \\
 & \alpha_y = s \\
 & v_y = f(x, y, u, \alpha, v, r, s) \\
 (*) \quad & p_y = v_y \\
 & r_y = s_x \\
 & s_y = f_x + f_u \cdot p + f_\alpha \cdot r + f_v \cdot s + f_r \cdot r_x + f_s \cdot s_x.
 \end{aligned}$$

Setting $\phi = (u, \alpha, v, p, r, s)$, we can write this system as

$$\phi_y = A\phi_x + \psi \quad \text{where} \quad A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & f_r & f_s & 0 \end{pmatrix}.$$

The eigenvalues of the matrix

$$A = \begin{pmatrix} 0 & 1 \\ f_r & f_s \end{pmatrix}$$

are the roots of $\lambda^2 - f_s\lambda - f_r = 0$, namely

$$\frac{f_s \pm \sqrt{f_s^2 + 4f_r}}{2}.$$

These roots are real and distinct, since $f_s^2 + 4f_r > 0$, and they are $\neq 0$, since $f_r \neq 0$. So A' is diagonalizable, and consequently A is. Theorem 10 then tells us that we can solve the system (*) with the appropriate initial conditions; hence we can solve the original equation (I) with the given hyperbolic initial conditions. The only slight problem is that Theorem 10 was stated only for non-singular A (although, as we mentioned, this requirement is not really necessary). We will therefore give another approach, based on the work of H. Lewy [4], which uses Theorem 10 only as stated, and we will work directly with equation (I), rather than (I'), without assuming that the initial curve is a segment of the x -axis. The new considerations which we will introduce are not only interesting in their own right, but will also be used in the next section.

We are thus considering the second order equation

$$(I) \quad 0 = F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = F(x, y, u, p, q, r, s, t),$$

together with an initial curve $c = (c_1, c_2): [a, b] \rightarrow \mathbb{R}^2$ along with initial data $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}, \overset{\circ}{r}, \overset{\circ}{s}, \overset{\circ}{t}: [a, b] \rightarrow \mathbb{R}$ satisfying

$$(I-1) \quad \begin{aligned} 0 &= F(c_1(\tau), c_2(\tau), \overset{\circ}{u}(\tau), \overset{\circ}{p}(\tau), \overset{\circ}{q}(\tau), \overset{\circ}{r}(\tau), \overset{\circ}{s}(\tau), \overset{\circ}{t}(\tau)) \\ &= F(C(\tau)), \quad \text{say.} \end{aligned}$$

In addition, the initial data must satisfy (compare page 34)

$$(I-2) \quad \begin{aligned} \frac{d\overset{\circ}{u}}{d\tau} &= \overset{\circ}{p} \frac{dc_1}{d\tau} + \overset{\circ}{q} \frac{dc_2}{d\tau} \\ \frac{d\overset{\circ}{p}}{d\tau} &= \overset{\circ}{r} \frac{dc_1}{d\tau} + \overset{\circ}{s} \frac{dc_2}{d\tau} \\ \frac{d\overset{\circ}{q}}{d\tau} &= \overset{\circ}{s} \frac{dc_1}{d\tau} + \overset{\circ}{t} \frac{dc_2}{d\tau}. \end{aligned}$$

The fact that the initial data is hyperbolic is expressed by the inequality

$$(I-3) \quad F_s^2 - 4F_r F_t > 0 \quad \text{for all } C(\tau).$$

In section 2 we wrote the condition that c be free as

$$F_r v_1^2 + F_s v_1 v_2 + F_t v_2^2 \neq 0$$

where (v_1, v_2) is a normal at $c(\tau)$ [and F_r, F_s, F_t are evaluated at $C(\tau)$]. If (X, Y) is tangent to c at τ , then the normal is a multiple of $(Y, -X)$, so we can also write this condition as

$$(I-4) \quad F_r Y^2 - F_s XY + F_t X^2 \neq 0.$$

By contrast, the characteristic directions are just those (X, Y) satisfying

$$(I-C) \quad F_r Y^2 - F_s XY + F_t X^2 = 0.$$

Arguments similar to the one given at the beginning of this section show that without loss of generality we can assume that we have $F_r, F_t \neq 0$ at all $C(\tau)$. Now in a neighborhood $\mathcal{U} \subset \mathbb{R}^8$ of $\{C(\tau)\}$ where we have

$$\begin{aligned} F_s^2 - 4F_r F_t &> 0 \\ F_r, F_t &\neq 0 \end{aligned}$$

we can find two continuous everywhere unequal real-valued functions ρ_1, ρ_2 which are solutions of

$$(a) \quad F_r \rho^2 - F_s \rho + F_t = 0.$$

The condition $F_r \neq 0$ guarantees that we have a genuine quadratic equation. The condition $F_t \neq 0$ insures that $\rho_1, \rho_2 \neq 0$, which will be important later on.

Now suppose that we actually have a solution u of (I). Let $\mathcal{V} \subset \mathbb{R}^2$ be $\{(x, y) : (x, y, u(x, y), u_x(x, y), \dots) \in \mathcal{U}\}$. For $R_i = \rho_i(x, y, u(x, y), \dots)$, the two vectors $(1, R_1)$ and $(1, R_2)$ are linearly independent at each point of \mathcal{V} [and equation (I-C) shows that they are always characteristic]. We can now use Proposition I.5-19 to choose an open set $\mathcal{W} \subset \mathbb{R}^2$, with standard coordinates (ξ, η) , say, and a diffeomorphism $\phi: \mathcal{W} \rightarrow \mathcal{V}$ such that

- (i) the parameter curves of ϕ are always characteristic, that is, their tangent vectors are multiples of $(1, R_1)$ or $(1, R_2)$,
- (ii) $\phi(\xi, \xi) = c(\xi)$ for $\xi \in [a, b]$.

Let \mathbf{x} and \mathbf{y} denote the compositions $x \circ \phi$ and $y \circ \phi$, for the standard coordinate system (x, y) on \mathcal{V} . Thus, \mathbf{x} and \mathbf{y} are just the component functions of ϕ . Similarly, let $\mathbf{u} = u \circ \phi$, $\mathbf{p} = p \circ \phi = u_x \circ \phi$, etc. For brevity, if $g: \mathcal{W} \rightarrow \mathbb{R}$ we will use g' to denote the partial derivative $\partial g / \partial \xi$, and g^{\setminus} to denote $\partial g / \partial \eta$. Condition (i) on our map ϕ means that

$$(*1) \quad \rho_1 \mathbf{x}' - \mathbf{y}' = 0$$

$$(*2) \quad \rho_2 \mathbf{x}^{\setminus} - \mathbf{y}^{\setminus} = 0,$$

where ρ_i actually means the function whose value at (ξ, η) is

$$\rho_i(\phi(\xi, \eta), u(\phi(\xi, \eta)), u_x(\phi(\xi, \eta)), \dots) = \rho_i(\mathbf{x}(\xi, \eta), \mathbf{y}(\xi, \eta), \mathbf{u}(\xi, \eta), \mathbf{p}(\xi, \eta), \dots).$$

We also have the general equations

$$(*3) \quad \mathbf{u}' - \mathbf{p}\mathbf{x}' - \mathbf{q}\mathbf{y}' = 0$$

$$(*4) \quad \mathbf{p}' - \mathbf{r}\mathbf{x}' - \mathbf{s}\mathbf{y}' = 0$$

$$(*5) \quad \mathbf{q}' - \mathbf{s}\mathbf{x}' - \mathbf{t}\mathbf{y}' = 0.$$

We obtain further equations as follows. When u is a solution of (I) we have

$$\begin{aligned} 0 &= \frac{\partial F(x, y, u(x, y), \dots)}{\partial x} \\ &= F_x + F_u \cdot u_x + F_p \cdot p_x + F_q \cdot q_x \\ &\quad + F_r \cdot r_x + F_s \cdot s_x + F_t \cdot t_x \\ &= F_r \cdot r_x + F_s \cdot s_x + F_t \cdot t_x + \dots, \end{aligned}$$

where F_x, F_u, F_p, \dots are evaluated at $(x, y, u(x, y), \dots)$. Composing with ϕ , we can write

$$(A) \quad 0 = F_r \cdot (r_x \circ \phi) + F_s \cdot (s_x \circ \phi) + F_t \cdot (t_x \circ \phi) + \{F\}_x,$$

where $\{F\}_x = F_x + F_u \cdot p + F_p \cdot q + F_q \cdot r,$

all partials of F being evaluated at (x, y, u, p, \dots) .

On the other hand, we always have

$$\begin{aligned} r' &= (r_x \circ \phi) \cdot x' + (r_y \circ \phi) \cdot y' \\ s' &= (s_x \circ \phi) \cdot x' + (s_y \circ \phi) \cdot y', \end{aligned}$$

and thus, using equality of mixed partials,

$$(B) \quad \begin{aligned} r' &= (r_x \circ \phi) \cdot x' + (s_x \circ \phi) \cdot y' \\ s' &= (s_x \circ \phi) \cdot x' + (t_x \circ \phi) \cdot y'. \end{aligned}$$

We can write equations (A) and (B) together in matrix form as

$$(C) \quad \begin{pmatrix} x' & y' & 0 \\ 0 & x' & y' \\ F_r & F_s & F_t \end{pmatrix} \cdot \begin{pmatrix} r_x \circ \phi \\ s_x \circ \phi \\ t_x \circ \phi \end{pmatrix} = \begin{pmatrix} r' \\ s' \\ -\{F\}_x \end{pmatrix}.$$

Now

$$\begin{aligned} \det \begin{pmatrix} x' & y' & 0 \\ 0 & x' & y' \\ F_r & F_s & F_t \end{pmatrix} &= (x')^2 F_t - x' y' F_s + (y')^2 F_r \\ &= (x')^2 (F_t - \rho_1 F_s + \rho_1^2 F_r) \quad \text{by } (*1) \\ &= 0, \quad \text{since } \rho_1 \text{ is a solution of (a).} \end{aligned}$$

Consequently

$$\text{rank} \begin{pmatrix} x' & y' & 0 \\ 0 & x' & y' \\ F_r & F_s & F_t \end{pmatrix} \leq 2.$$

Moreover, the last column of the matrix

$$\begin{pmatrix} x' & y' & 0 & r' \\ 0 & x' & y' & s' \\ F_r & F_s & F_t & -\{F\}_x \end{pmatrix}$$

is a linear combination of the first three columns, namely the linear combination $(r_x \circ \phi, s_x \circ \phi, t_x \circ \phi)$, by equation (C), so this matrix also has rank ≤ 2 . Consequently, the determinant of every 3×3 submatrix vanishes. In particular,

$$0 = \det \begin{pmatrix} x' & 0 & r' \\ 0 & y' & s' \\ F_r & F_t & -\{F\}_x \end{pmatrix} = -y' F_r r' - x' F_t s' - x' \{F\}_x y';$$

using (*1) this can be written as follows (note that we don't substitute for the second y'):

$$(*6) \quad \rho_1 F_r r' + F_t s' + \{F\}_x y' = 0.$$

Similarly, we have the following equation, which we will number as

$$(*8) \quad \rho_2 F_r r' + F_t s' + \{F\}_x y' = 0$$

Exactly the same manipulations may be carried out by differentiating with respect to y , instead of x . We have the equation

$$0 = F_r \cdot (r_y \circ \phi) + F_s \cdot (s_y \circ \phi) + F_t \cdot (t_y \circ \phi) + \{F\}_y,$$

together with

$$\begin{aligned} s' &= (r_y \circ \phi) \cdot x' + (s_y \circ \phi) \cdot y' \\ t' &= (s_y \circ \phi) \cdot x' + (t_y \circ \phi) \cdot y', \end{aligned}$$

which we can write as

$$\begin{pmatrix} x' & y' & 0 \\ 0 & x' & y' \\ F_r & F_s & F_t \end{pmatrix} \cdot \begin{pmatrix} r_y \circ \phi \\ s_y \circ \phi \\ t_y \circ \phi \end{pmatrix} = \begin{pmatrix} s' \\ t' \\ -\{F\}_y \end{pmatrix}.$$

So, as before, we have

$$(*7) \quad \rho_1 F_r s' + F_t t' + \{F\}_y y' = 0$$

(as well as an equation involving \cdot).

We have thus selected 8 equations satisfied by the 8 functions x, y, u, p, q, r, s, t when u is a solution of (I):

$$\begin{aligned}
 (*1) \quad & \rho_1 x' - y' = 0 \\
 (*2) \quad & \rho_2 x' - y' = 0 \\
 (*3) \quad & u' - px' - qy' = 0 \\
 (*4) \quad & p' - rx' - sy' = 0 \\
 (*5) \quad & q' - sx' - ty' = 0 \\
 (*6) \quad & \rho_1 F_r r' + F_t s' + (F_x + F_u p + F_p r + F_q s) y' = 0 \\
 (*7) \quad & \rho_1 F_r s' + F_t t' + (F_y + F_u q + F_p s + F_q t) y' = 0 \\
 (*8) \quad & \rho_2 F_r r' + F_t s' + (F_x + F_u p + F_p r + F_q s) y' = 0.
 \end{aligned}$$

We have

$$x(\xi, \xi) = x(\phi(\xi, \xi)) = x(c(\xi)) = c_1(\xi),$$

and similarly for y . And if u has the initial data for (I), then

$$u(\xi, \xi) = u(\phi(\xi, \xi)) = u(c(\xi)) = \overset{\circ}{u}(\xi),$$

and similarly for p, q, \dots . Thus, along the line segment $S = \{(\xi, \xi) : a \leq \xi \leq b\}$ the solutions x, y, u, \dots of (*1)–(*8) have the initial conditions

$$(*0) \quad \begin{cases} x(\xi, \xi) = c_1(\xi), & y(\xi, \xi) = c_2(\xi) \\ u(\xi, \xi) = \overset{\circ}{u}(\xi) \\ p(\xi, \xi) = \overset{\circ}{p}(\xi), & q(\xi, \xi) = \overset{\circ}{q}(\xi) \\ r(\xi, \xi) = \overset{\circ}{r}(\xi), & s(\xi, \xi) = \overset{\circ}{s}(\xi), \quad t(\xi, \xi) = \overset{\circ}{t}(\xi). \end{cases}$$

The beauty of this particular set of 8 equations is the fact that they automatically lead to solutions of equation (I). More precisely, consider 8 functions x, y, u, \dots, t satisfying the system (*), with the initial conditions (*0). We are denoting the 8 unknowns of our system by x, y, u, \dots, t simply for convenience—we are *not* assuming that $D_1 u = p$, etc.

11. LEMMA. Let $c, \overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}, \overset{\circ}{r}, \overset{\circ}{s}, \overset{\circ}{t} : [a, b] \rightarrow \mathbb{R}$ satisfy (I-1)–(I-4). Suppose that x, y, u, p, q, r, s, t satisfy the system (*), with the initial conditions (*0). Then (x, y) is a coordinate system in a neighborhood of the diagonal line segment $S = \{(\xi, \xi) : a \leq \xi \leq b\}$, and $u \circ (x, y)^{-1}$ is a solution of the PDE (I) with the initial conditions $\overset{\circ}{u}, \overset{\circ}{p}, \overset{\circ}{q}, \overset{\circ}{r}, \overset{\circ}{s}, \overset{\circ}{t}$ on the curve c .

PROOF. Equations (*1) and (*2) show that $(\mathbf{x}', \mathbf{y}')$ and $(\mathbf{x}', \mathbf{y}')$ are characteristic directions, while c' is never characteristic by (I-4). In particular, $c'(\xi)$ is not a multiple of $(\mathbf{x}', \mathbf{y}')(\xi, \xi)$ or $(\mathbf{x}', \mathbf{y}')(\xi, \xi)$.

Our initial conditions on \mathbf{x} and \mathbf{y} give

$$\begin{aligned}\mathbf{x}'(\xi, \xi) + \mathbf{x}'(\xi, \xi) &= c_1'(\xi) \\ \mathbf{y}'(\xi, \xi) + \mathbf{y}'(\xi, \xi) &= c_2'(\xi).\end{aligned}$$

If we had $\mathbf{x}'(\xi, \xi) = 0$ then we would also have $\mathbf{y}'(\xi, \xi) = 0$ by (*1), which would make $(\mathbf{x}', \mathbf{y}')(\xi, \xi) = c'(\xi)$, which we have just noted is not possible. Similarly, we cannot have $\mathbf{x}'(\xi, \xi) = 0$. Thus, the map

$$\phi(\xi, \eta) = (\mathbf{x}(\xi, \eta), \mathbf{y}(\xi, \eta))$$

has Jacobian matrix

$$\begin{pmatrix} \mathbf{x}' & \mathbf{y}' \\ \mathbf{x}' & \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{x}' & \rho_1 \mathbf{x}' \\ \mathbf{x}' & \rho_2 \mathbf{x}' \end{pmatrix}$$

with determinant

$$\mathbf{x}' \mathbf{x}' \cdot (\rho_2 - \rho_1) \neq 0 \quad \text{at all points of } S.$$

So $\phi = (\mathbf{x}, \mathbf{y})$ is a coordinate system in a neighborhood of the compact set S .

We next claim that we have

$$(1) \quad 0 = F(\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{s}, \mathbf{t}) \quad [= F(\mathbf{x}(\xi, \eta), \mathbf{y}(\xi, \eta), \mathbf{u}(\xi, \eta), \dots)].$$

To prove this, we add ρ_1 times equation (*7) to (*6), and replace \mathbf{y}' in equation (*6) by $\rho_1 \mathbf{x}'$ [using (*1)]; we thus obtain

$$\begin{aligned}\rho_1^2 F_r s' + \rho_1 F_t t' + \rho_1 (F_y + F_u \mathbf{q} + F_p \mathbf{s} + F_q \mathbf{t}) \mathbf{y}' \\ + \rho_1 F_r \mathbf{r}' + F_t s' + \rho_1 (F_x + F_u \mathbf{p} + F_p \mathbf{r} + F_q \mathbf{s}) \mathbf{x}' = 0.\end{aligned}$$

The coefficient of s' is

$$\rho_1^2 F_r + F_t = \rho_1 F_s,$$

so after dividing by $\rho_1 \neq 0$ we obtain

$$\begin{aligned}0 = F_r \mathbf{r}' + F_s s' + F_t t' + (F_x + F_u \mathbf{p} + F_p \mathbf{r} + F_q \mathbf{s}) \mathbf{x}' \\ + (F_y + F_u \mathbf{q} + F_p \mathbf{s} + F_q \mathbf{t}) \mathbf{y}'.\end{aligned}$$

Making use of (*3)(*5) we then have

$$\begin{aligned} 0 &= F_r r' + F_s s' + F_t t' + F_x x' + F_y y' + F_u u' + F_p p' + F_q q' \\ &= F'. \end{aligned}$$

On the other hand, since $\hat{u}, \hat{p}, \hat{q}, \dots$ are assumed to satisfy (I-2), the initial conditions (*₀) insure that $F = 0$ on the diagonal line segment S . Therefore we have $F = 0$ in a whole neighborhood of this interval.

To prove that $u \circ \phi^{-1} = u \circ (x, y)^{-1}$ is a solution to the equation (I) it thus suffices to prove that

$$p \circ \phi^{-1} = D_1(u \circ \phi^{-1}), \quad q \circ \phi^{-1} = D_2(u \circ \phi^{-1})$$

$$r \circ \phi^{-1} = D_1(p \circ \phi^{-1}), \quad s \circ \phi^{-1} = D_2(p \circ \phi^{-1}), \quad t \circ \phi^{-1} = D_2(q \circ \phi^{-1}).$$

Note that, with our standard notation $\partial/\partial x$ and $\partial/\partial y$ for the coordinate system (x, y) we have

$$D_1(u \circ \phi^{-1}) = \frac{\partial u}{\partial x} \circ \phi^{-1}, \text{ etc.,}$$

so what we have to show amounts to

$$\begin{aligned} p &= \frac{\partial u}{\partial x}, & q &= \frac{\partial u}{\partial y} \\ r &= \frac{\partial p}{\partial x}, & s &= \frac{\partial p}{\partial y}, & t &= \frac{\partial q}{\partial y}. \end{aligned}$$

By our usual chain rule, we have, for any function α ,

$$\begin{aligned} \frac{\partial \alpha}{\partial x} &= \frac{\partial \alpha}{\partial \xi} \cdot \frac{\partial \xi}{\partial x} + \frac{\partial \alpha}{\partial \eta} \cdot \frac{\partial \eta}{\partial x} \\ \frac{\partial \alpha}{\partial y} &= \frac{\partial \alpha}{\partial \xi} \cdot \frac{\partial \xi}{\partial y} + \frac{\partial \alpha}{\partial \eta} \cdot \frac{\partial \eta}{\partial y}, \end{aligned}$$

or, with our abbreviations,

$$\begin{aligned} \frac{\partial \alpha}{\partial x} &= \alpha' \cdot \frac{\partial \xi}{\partial x} + \alpha'' \cdot \frac{\partial \eta}{\partial x} \\ \frac{\partial \alpha}{\partial y} &= \alpha' \cdot \frac{\partial \xi}{\partial y} + \alpha'' \cdot \frac{\partial \eta}{\partial y}. \end{aligned}$$

The partials $\partial\xi/\partial\mathbf{x}, \dots$ are given by

$$\begin{pmatrix} \partial\xi/\partial\mathbf{x} & \partial\xi/\partial\mathbf{y} \\ \partial\eta/\partial\mathbf{x} & \partial\eta/\partial\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{x}' & \mathbf{x}' \\ \mathbf{y}' & \mathbf{y}' \end{pmatrix}^{-1} = \frac{1}{\mathbf{x}'\mathbf{y}' - \mathbf{x}'\mathbf{y}'} \begin{pmatrix} \mathbf{y}' & -\mathbf{x}' \\ -\mathbf{y}' & \mathbf{x}' \end{pmatrix}.$$

Thus

$$\frac{\partial\alpha}{\partial\mathbf{x}} = \frac{\alpha'\mathbf{y}' - \alpha'\mathbf{y}'}{\mathbf{x}'\mathbf{y}' - \mathbf{x}'\mathbf{y}'}, \quad \frac{\partial\alpha}{\partial\mathbf{y}} = \frac{-\alpha'\mathbf{x}' + \alpha'\mathbf{x}'}{\mathbf{x}'\mathbf{y}' - \mathbf{s}'\mathbf{y}'}.$$

which can be written as

$$\begin{aligned} (2) \quad \mathbf{x}'\alpha' - \mathbf{x}'\alpha' &= \frac{\partial\alpha}{\partial\mathbf{y}}(\mathbf{x}'\mathbf{y}' - \mathbf{x}'\mathbf{y}') \\ \mathbf{y}'\alpha' - \mathbf{y}'\alpha' &= \frac{\partial\alpha}{\partial\mathbf{x}}(\mathbf{y}'\mathbf{x}' - \mathbf{y}'\mathbf{x}'). \end{aligned}$$

Now equations (*6) and (*8) can be written in the form

$$\begin{aligned} \frac{F_r}{\mathbf{x}'}\mathbf{r}' + \frac{F_t}{\mathbf{y}'}\mathbf{s}' + F_x + F_u\mathbf{p} + F_p\mathbf{r} + F_q\mathbf{s} &= 0 \\ \frac{F_r}{\mathbf{x}'}\mathbf{r}' + \frac{F_t}{\mathbf{y}'}\mathbf{s}' + F_x + F_u\mathbf{p} + F_p\mathbf{r} + F_q\mathbf{s} &= 0. \end{aligned}$$

Hence

$$\frac{F_r}{\mathbf{x}'}\mathbf{r}' + \frac{F_t}{\mathbf{y}'}\mathbf{s}' = \frac{F_r}{\mathbf{x}'}\mathbf{r}' + \frac{F_t}{\mathbf{y}'}\mathbf{s}'.$$

Using (2), this equation can be written

$$(3) \quad F_r \frac{\partial\mathbf{r}}{\partial\mathbf{y}} \left(\frac{\mathbf{y}'}{\mathbf{x}'} - \frac{\mathbf{y}'}{\mathbf{x}'} \right) + F_t \frac{\partial\mathbf{s}}{\partial\mathbf{x}} \left(\frac{\mathbf{x}'}{\mathbf{y}'} - \frac{\mathbf{x}'}{\mathbf{y}'} \right) = 0.$$

But

$$\frac{\mathbf{x}'}{\mathbf{y}'} \cdot \frac{\mathbf{x}'}{\mathbf{y}'} = \frac{1}{\rho_1\rho_2} = \frac{F_r}{F_t},$$

so from (3) we obtain the preliminary result

$$(4) \quad \frac{\partial\mathbf{r}}{\partial\mathbf{y}} = \frac{\partial\mathbf{s}}{\partial\mathbf{x}}.$$

Next note that we have

$$\begin{aligned} \mathbf{p}' &= \frac{\partial\mathbf{p}}{\partial\mathbf{x}}\mathbf{x}' + \frac{\partial\mathbf{p}}{\partial\mathbf{y}}\mathbf{y}' && \text{by the chain rule} \\ \mathbf{p}' &= \mathbf{r}\mathbf{x}' + \mathbf{s}\mathbf{y}' && \text{by (*4),} \end{aligned}$$

so that

$$(5) \quad \mathbf{x}' \left(\mathbf{r} - \frac{\partial \mathbf{p}}{\partial \mathbf{x}} \right) = \mathbf{y}' \left(\mathbf{s} - \frac{\partial \mathbf{p}}{\partial \mathbf{y}} \right).$$

On the other hand, the initial conditions $(*_0)$ give

$$\begin{aligned} \frac{d \circ \bar{\mathbf{p}}}{d\xi} &= \frac{d}{d\xi} \mathbf{p}(\xi, \xi) \\ &= \frac{\partial \mathbf{p}}{\partial \mathbf{x}}(\xi, \xi) \frac{d}{d\xi} \mathbf{x}(\xi, \xi) + \frac{\partial \mathbf{p}}{\partial \mathbf{y}}(\xi, \xi) \frac{d}{d\xi} \mathbf{y}(\xi, \xi) \\ &= \frac{\partial \mathbf{p}}{\partial \mathbf{x}}(\xi, \xi) c_1'(\xi) + \frac{\partial \mathbf{p}}{\partial \mathbf{y}}(\xi, \xi) c_2'(\xi), \end{aligned}$$

while the conditions (I-2) imply that

$$\frac{d \circ \bar{\mathbf{p}}}{d\xi} = \mathbf{r}(\xi, \xi) c_1'(\xi) + \mathbf{s}(\xi, \xi) c_2'(\xi),$$

so that

$$(6) \quad c_1'(\xi) \cdot \left[\mathbf{r} - \frac{\partial \mathbf{p}}{\partial \mathbf{x}} \right](\xi, \xi) = c_2'(\xi) \cdot \left[\mathbf{s} - \frac{\partial \mathbf{p}}{\partial \mathbf{y}} \right](\xi, \xi).$$

Equations (5) and (6) together give

$$\begin{pmatrix} c_1'(\xi) & c_2'(\xi) \\ \mathbf{x}'(\xi, \xi) & \mathbf{y}'(\xi, \xi) \end{pmatrix} \begin{pmatrix} \left[\mathbf{r} - \frac{\partial \mathbf{p}}{\partial \mathbf{x}} \right](\xi, \xi) \\ \left[\mathbf{s} - \frac{\partial \mathbf{p}}{\partial \mathbf{y}} \right](\xi, \xi) \end{pmatrix} = 0.$$

Since c' is not a multiple of $(\mathbf{x}', \mathbf{y}')$, as we observed at the beginning of the proof, it follows that we must have

$$\mathbf{r}(\xi, \xi) = \frac{\partial \mathbf{p}}{\partial \mathbf{x}}(\xi, \xi), \quad \mathbf{s}(\xi, \xi) = \frac{\partial \mathbf{p}}{\partial \mathbf{y}}(\xi, \xi).$$

So, if we set

$$P = \mathbf{p}' - \mathbf{r}\mathbf{x}' - \mathbf{s}\mathbf{y}',$$

then $P(\xi, \xi) = 0$. Moreover,

$$P' = \mathbf{p}'' - \mathbf{r}'\mathbf{x}' - \mathbf{s}'\mathbf{y}' - \mathbf{r}\mathbf{x}'' - \mathbf{s}\mathbf{y}'',$$

while equation (*4) gives

$$0 = p'^{\wedge} - r'x' - s'y' - rx'^{\wedge} - sy'^{\wedge}.$$

So we have

$$\begin{aligned} P' &= r'x' - r'x' + s'y' - s'y' \\ &= \left(\frac{\partial s}{\partial x} - \frac{\partial r}{\partial y} \right) (x'y' - x'y') \quad \text{by (2)} \\ &= 0 \quad \text{by (4).} \end{aligned}$$

Consequently, we have

$$(7) \quad 0 = P = p' - rx' - sy'$$

in a neighborhood of S . From (*4) and (7) we have

$$(8) \quad \begin{aligned} x'p' - x'p' &= s(x'y' - x'y') \\ y'p' - y'p' &= r(y'x' - y'x'). \end{aligned}$$

So (2) gives

$$(9) \quad r = \frac{\partial p}{\partial x}, \quad s = \frac{\partial p}{\partial y}.$$

Similarly, set

$$\begin{aligned} U &= u' - px' - qy' \\ Q &= q' - sx' - ty'. \end{aligned}$$

Then we have

$$U = u'^{\wedge} - p'x' - q'y' - px'^{\wedge} - qy'^{\wedge},$$

and

$$0 = u'^{\wedge} - p'x' - q'y' - px'^{\wedge} - qy'^{\wedge} \quad \text{from (*3),}$$

and thus

$$\begin{aligned} (10) \quad U' &= -p'x' - q'y' + p'x' + q'y' \\ &= -s(x'y' - x'y') - q'y' + q'y' \quad \text{by (8)} \\ &= (q' - sx' - ty')y' - (q' - sx' - ty')y' \\ &= Qy' \quad \text{by (*5).} \end{aligned}$$

Likewise,

$$\begin{aligned} Q' &= -s'x' - t'y' + s'x' + t'y' \\ &= \left(\frac{\partial t}{\partial x} - \frac{\partial s}{\partial y} \right) (x'y' - x'y') \quad \text{by (2).} \end{aligned}$$

Now (1) implies, using (9), that

$$0 = \frac{\partial F}{\partial x} = F_x + F_u \frac{\partial u}{\partial x} + F_p r + F_q \frac{\partial q}{\partial x} + F_r \frac{\partial r}{\partial x} + F_s \frac{\partial s}{\partial x} + F_t \frac{\partial t}{\partial x},$$

with a similar equation for $\partial F/\partial y$. Subtract the first of these equations from (*6) and the second from (*7), multiply the resulting equations by x' and y' , respectively, and add them. Taking into account (4), we obtain finally

$$\begin{aligned} &F_t \left(\frac{\partial t}{\partial x} - \frac{\partial s}{\partial y} \right) x' + F_q \left(\frac{\partial q}{\partial x} - s \right) x' + F_u \left(\frac{\partial u}{\partial x} - p \right) x' \\ &\quad - F_t \frac{x'}{y'} \left(\frac{\partial t}{\partial x} - \frac{\partial s}{\partial y} \right) y' + F_q \left(\frac{\partial q}{\partial y} - t \right) y' + F_u \left(\frac{\partial u}{\partial y} - q \right) y' = 0. \end{aligned}$$

Thus we have

$$\begin{aligned} \frac{F_t}{y'} Q' + F_q Q + F_u U &= 0 \\ U' &= Q y'. \end{aligned}$$

Along each line parallel to the ξ -axis, this is simply a system of ordinary differential equations for U and Q . Since $U = Q = 0$ on the diagonal segment S , we find that $U = Q = 0$ in a neighborhood of S . As before,

$$\begin{aligned} U = 0 \text{ and } (*3) &\implies p = \frac{\partial u}{\partial x}, & q = \frac{\partial u}{\partial y}, \\ Q = 0 \text{ and } (*5) &\implies s = \frac{\partial q}{\partial x}, & t = \frac{\partial q}{\partial y}, \end{aligned}$$

and this completes the proof. ♦

Now we want to know whether the system (*) does in fact have a solution with the initial conditions (*₀). We introduce the rotated coordinates (μ, ν) by

$$\begin{aligned}\mu &= \xi + \eta \\ \nu &= \xi - \eta.\end{aligned}$$

Then the segment $S = \{(\xi, \xi) : a \leq \xi \leq b\}$ corresponds to the segment $[2a, 2b]$ of the μ -axis, and

$$\begin{aligned}' &= \frac{\partial}{\partial \mu} + \frac{\partial}{\partial \nu} \\ ' &= \frac{\partial}{\partial \mu} - \frac{\partial}{\partial \nu}.\end{aligned}$$

Denoting x, y, u, \dots, t by ϕ_1, \dots, ϕ_8 , our equations (*1)–(*8) can be written as the following matrix equation, in which each row of the matrix \mathcal{Q} on the right is simply -1 times the corresponding row of the matrix on the left, except for rows 2 and 8, which are equal to $+1$ times the corresponding row [these rows correspond to equations (*2) and (*8) involving $'$]:

$$\begin{bmatrix} \rho_1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -p & -q & 1 & 0 & 0 & 0 & 0 & 0 \\ -r & -s & 0 & 1 & 0 & 0 & 0 & 0 \\ -s & -t & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \{F\}_x & 0 & 0 & 0 & \rho_1 F_r & F_t & 0 \\ 0 & \{F\}_y & 0 & 0 & 0 & 0 & \rho_1 F_r & F_t \\ 0 & \{F\}_x & 0 & 0 & 0 & \rho_2 F_r & F_t & 0 \end{bmatrix} \cdot \begin{bmatrix} \partial \phi_1 / \partial \mu \\ \partial \phi_2 / \partial \mu \\ \partial \phi_3 / \partial \mu \\ \partial \phi_4 / \partial \mu \\ \partial \phi_5 / \partial \mu \\ \partial \phi_6 / \partial \mu \\ \partial \phi_7 / \partial \mu \\ \partial \phi_8 / \partial \mu \end{bmatrix} = \mathcal{Q} \cdot \begin{bmatrix} \partial \phi_1 / \partial \nu \\ \partial \phi_2 / \partial \nu \\ \partial \phi_3 / \partial \nu \\ \partial \phi_4 / \partial \nu \\ \partial \phi_5 / \partial \nu \\ \partial \phi_6 / \partial \nu \\ \partial \phi_7 / \partial \nu \\ \partial \phi_8 / \partial \nu \end{bmatrix}.$$

The determinant of the matrix on the left is easily seen to be

$$(\rho_2 - \rho_1) \det \begin{pmatrix} \rho_1 F_r & F_t & 0 \\ 0 & \rho_1 F_r & F_t \\ \rho_2 F_r & F_t & 0 \end{pmatrix} = (\rho_2 - \rho_1)^2 F_r F_t^2 \neq 0.$$

Writing our equation for short as

$$\mathcal{P} \cdot \phi_\mu = \mathcal{Q} \cdot \phi_\nu,$$

we have

$$\mathcal{Q} = \mathcal{D} \mathcal{P},$$

where \mathcal{D} is the diagonal matrix with -1 's on all diagonals except for $+1$'s at

positions (2, 2) and (8, 8). So we can write our equation as

$$\begin{aligned}\phi_v &= \mathcal{Q}^{-1} \mathcal{P} \phi_\mu \\ &= \mathcal{P}^{-1} \mathcal{D}^{-1} \mathcal{P} \phi_\mu \\ &= \mathcal{P}^{-1} \mathcal{D} \mathcal{P} \phi_\mu.\end{aligned}$$

Since $\mathcal{P}^{-1} \mathcal{D} \mathcal{P}$ is diagonalizable, Theorem 10 shows that we can solve the system (*) uniquely with the initial conditions (*₀). So Lemma 11 shows that our original PDE (I) has a solution with the given initial conditions. We summarize this result in

12. THEOREM. Consider a second order equation

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0,$$

an initial curve, and hyperbolic initial data along this curve satisfying (I-1) and (I-2), such that the curve is free for the initial data. Suppose that the initial curve, initial data, and F have continuous third partial derivatives satisfying a Lipschitz condition. Then in a neighborhood of this initial curve there is a unique solution with this initial data having continuous second partial derivatives satisfying a Lipschitz condition.

Actually, the hypotheses could be weakened considerably here (partly by weakening the hypotheses in Theorem 10, and even more so by using other methods), but we will not worry about this. More important, we would like to know that the solution u has a high order of differentiability if F , the initial curve, and the initial data do, but we must be content with merely asserting this, since we did not prove this for Theorem 10. One thing is clear, however. Even if F and the initial curve are highly differentiable, or even analytic, there may be solutions u which are far less differentiable. In fact, if we have any hyperbolic initial data, then nearby initial data will also be hyperbolic. We can choose this nearby data to be only C^3 , and then our solution is at most C^3 . As we will see in the next section, this is in marked contrast to the situation for elliptic solutions.

We briefly indicate the situation for the case where our initial curve is everywhere characteristic for the initial data. Assuming, as at the beginning of this section, that the initial curve is a segment $[a, b]$ of the x -axis, the initial data amount to functions

$$\overset{\circ}{u}, \overset{\circ}{q}, \overset{\circ}{t}$$

on $[a, b]$ satisfying

$$(I_0) \quad 0 = F(x, 0, \overset{\circ}{u}(x), \overset{\circ}{u}'(x), \overset{\circ}{q}(x), \overset{\circ}{u}''(x), \overset{\circ}{q}'(x), \overset{\circ}{t}(x)),$$

except that now we have

$$(I_C) \quad 0 = F_t(x, 0, \overset{\circ}{u}(x), \overset{\circ}{u}'(x), \overset{\circ}{q}(x), \overset{\circ}{u}''(x), \overset{\circ}{q}'(x), \overset{\circ}{t}(x)).$$

Differentiating (I_0) and using (I_C) , we find that we must have

$$(**) \quad F_r \overset{\circ}{u}''' + F_s \overset{\circ}{q}'' + (F_x + F_u \overset{\circ}{u}' + F_p \overset{\circ}{u}'' + F_q \overset{\circ}{q}') = 0,$$

and conversely, if $(**)$ and (I_C) are satisfied, and (I_0) holds at one point $(x_0, 0)$, then it holds everywhere on $[a, b]$. By choosing a free curve γ and initial data such that $\gamma(0) = (x_0, 0)$ and the initial data for γ agrees with our initial data along $[a, b]$ at the point $(x_0, 0)$, we can then find a solution having this initial data everywhere on the characteristic curve $[a, b]$.

One particular case will be very important in Chapter 12. We consider an equation

$$(1) \quad 0 = F(x, y, pq, r, s, t) = A(rt - s^2) + Br + Cs + Dt + E \\ = (Ar + D)t + (Br + Cs + E - As^2),$$

where A, \dots, E depend only on x, y, p, q , so that our equation is linear in r, s, t , and $rt - s^2$. Such equations are called "Monge-Ampère equations", and they are the kind we will always encounter. A calculation, using (VI) on page 50, shows that our equation remains a Monge-Ampère equation when we compose with a diffeomorphism of the plane.

Suppose that along the x -axis we have picked $\overset{\circ}{u}$ and $\overset{\circ}{q}$ as the first two functions for our initial data. We can already see if the x -axis is characteristic, because the condition for this is

$$(2) \quad 0 = F_t = Ar + D,$$

which only involves $\overset{\circ}{u}, \overset{\circ}{p} = \overset{\circ}{u}'$, and $\overset{\circ}{q}$ on the x -axis. If equation (2) holds, then there is no hope of selecting $\overset{\circ}{t}$ to complete our initial data unless we also have

$$(3) \quad 0 = Br - Cs + E - As^2$$

along the x -axis. On the other hand, if this equation holds, then any choice of $\overset{\circ}{t}$ will make $\overset{\circ}{u}, \overset{\circ}{q}, \overset{\circ}{t}$ satisfy (1), so if $\overset{\circ}{u}$ and $\overset{\circ}{q}$ satisfy (2) and (3), we can find solutions of (1) with initial data $\overset{\circ}{u}, \overset{\circ}{q}, \overset{\circ}{t}$ for arbitrary functions $\overset{\circ}{t}$.

9. ELLIPTIC SOLUTIONS OF SECOND ORDER EQUATIONS IN TWO VARIABLES

The most important topic in the study of elliptic equations is the Dirichlet problem, and if this were a book on PDE's, it would be inexcusable not to devote a great deal of time to this subject. But we won't say another word about it. Instead, we will consider another aspect of elliptic equations, which often isn't even mentioned in a first course in PDE's. We have already noted that every solution of $u_{xx} + u_{yy} = 0$ is automatically analytic. In this section we will prove that every elliptic solution of any second order equation $F(x, y, u, p, q, r, s, t) = 0$ is likewise analytic, provided of course that F is an analytic function of its arguments. This theorem holds for elliptic solutions of second order PDE's in any number of variables, but the proof we will give works only in the two variable case. This deficiency (which doesn't bother us, since we are interested only in the two variable case) is more than compensated for by its conceptual simplicity. Moreover, the proof, from H. Lewy [5], has one truly beautiful feature—there isn't a single inequality in it.

Since the details of the proof become somewhat complicated, it will probably help to first examine a special case. Consider the equation

$$u_{xx} + u_{yy} = f(x, y, u, p, q),$$

where f is a real analytic function of its arguments. We will show that u can be extended to a complex analytic function from \mathbb{C}^2 to \mathbb{C} ; consequently u must be real analytic.

We recall first that if

$$\alpha(x, y) = \beta(x, y) + i\gamma(x, y)$$

for real-valued β, γ , then the Cauchy-Riemann equations for α are

$$\beta_x = \gamma_y, \quad \beta_y = -\gamma_x;$$

these two equations are equivalent to the single equation

$$(1) \quad \alpha_x = -i\alpha_y.$$

We will rewrite our equation for u as

$$(2) \quad u_{x_1 x_1} + u_{y_1 y_1} = f(x_1, y_1, u, p, q).$$

We denote the two coordinates in \mathbb{R}^2 by x_1, y_1 so that we can consider $\mathbb{R}^2 \subset \mathbb{C}^2$, where \mathbb{C}^2 has coordinates $x = x_1 + ix_2, y = y_1 + iy_2$. Thus we think of u as

a function such that $u(x_1, 0, y_1, 0)$ is defined. We first want to find a complex-valued extension $u(x_1, 0, y_1, y_2)$ of u which is complex analytic in $y_1 + iy_2$. Equation (1) shows that our desired extension should satisfy

$$(3) \quad u_{x_1 x_1} - u_{y_2 y_2} = f(x_1, y_1 + iy_2, u, u_{x_1}, -iu_{y_2})$$

at all points $(x_1, 0, y_1, y_2)$. For each fixed y_1 , consider equation (3) in the (x_1, y_2) -plane, with the initial conditions

$$(4) \quad u(x_1, 0, y_1, 0) = \text{the original } u(x_1, 0, y_1, 0)$$

$$(5) \quad u_{y_2}(x_1, 0, y_1, 0) = i \cdot u_{y_1}(x_1, 0, y_1, 0), \quad \text{for the original } u(x_1, 0, y_1, 0).$$

Equation (3) is hyperbolic. Actually this statement is misleading, for the right side of (3) is already complex-valued, so we have to allow the solution u to be complex-valued; thus we have to consider (3) as an equation for the real and imaginary parts of u . However, if we replace (3) by a system of quasi-linear equations, and make a new system by looking at the real and imaginary parts of all the functions in the old system, then the new system will in fact be hyperbolic. The reader may check this (we will write things out explicitly later on, for the general case). Then Theorem 10 shows* that we really can solve (3) with initial conditions (4) and (5).

Differentiating (5) gives

$$u_{y_2 y_1} = i u_{y_1 y_1} \quad \text{at } (x_1, 0, y_1, 0),$$

while subtracting (3) from (2) gives

$$u_{y_1 y_1} + u_{y_2 y_2} = 0 \quad \text{at } (x_1, 0, y_1, 0).$$

From these two equations we have

$$(6) \quad u_{y_2 y_2} = i u_{y_1 y_2} \quad \text{at } (x_1, 0, y_1, 0).$$

Equations (5) and (6) can also be written

$$(7) \quad \begin{cases} u_{y_1} + i u_{y_2} = 0 \\ \frac{\partial}{\partial y_2}(u_{y_1} + i u_{y_2}) = 0 \end{cases} \quad \text{at } (x_1, 0, y_1, 0).$$

*If we use the system of equations derived in the previous section, then Theorem 10 suffices. If we use the system of equations derived in section 3, then we would need the stronger form of Theorem 10 which allows the matrix A to be singular.

On the other hand, we can also obtain an equation for $\omega = u_{y_1} + iu_{y_2}$. To equation (3) we apply the operator $\nabla = \partial/\partial y_1 + i\partial/\partial y_2$; in the notation of Addendum 1 to Chapter 9, $\frac{1}{2}\nabla u$ would be written $u_{\bar{y}}$. Then we have

$$\begin{aligned}\omega_{x_1 x_1} - \omega_{y_2 y_2} &= \nabla(u_{x_1 x_1} - u_{y_2 y_2}) \\ &= f_y \nabla y + f_u \nabla u + f_p \nabla(u_{x_1}) - i f_q \nabla(u_{y_2}) \\ &\quad \text{since } f \text{ is analytic (compare pg. IV.320)} \\ &= 0 + f_u \omega + f_p \omega_{x_1} - i f_q \omega_{y_2}.\end{aligned}$$

This is a hyperbolic system for ω . Thus (7) implies that $\omega = 0$, by uniqueness of solutions. Hence

$$(8) \quad u_{y_1}(x_1, 0, y_1, y_2) + iu_{y_2}(x_1, 0, y_1, y_2) = 0.$$

Now we will extend u to \mathbb{R}^4 . We do this by considering the equation

$$(9) \quad u_{y_1 y_1} - u_{x_2 x_2} = f(x_1 + ix_2, y, u, -iu_{x_2}, u_{y_2});$$

here x_1 and y_2 are the parameters. We use the initial conditions

$$(10) \quad u(x_1, 0, y_1, y_2) = \text{the } u(x_1, 0, y_1, y_2) \text{ already obtained}$$

$$(11) \quad u_{x_2}(x_1, 0, y_1, y_2) = i \cdot u_{x_1}(x_1, 0, y_1, y_2),$$

for the $u(x_1, 0, y_1, y_2)$ already obtained.

Again we obtain a hyperbolic system, so we can solve (9), with the initial conditions (10) and (11).

Differentiating (11) gives

$$u_{x_1 x_2} = iu_{x_1 x_1} \quad \text{at } (x_1, 0, y_1, y_2),$$

while differentiating (8) with respect to y_1 and y_2 , and then subtracting, gives

$$u_{y_1 y_1} + u_{y_2 y_2} = 0 \quad \text{at } (x_1, 0, y_1, y_2).$$

Finally, subtracting (3) from (9) gives

$$u_{y_1 y_1} + u_{y_2 y_2} - u_{x_1 x_1} - u_{x_2 x_2} = 0 \quad \text{at } (x_1, 0, y_1, y_2).$$

From these three equations we obtain

$$(12) \quad u_{x_2 x_2} = iu_{x_1 x_2} \quad \text{at } (x_1, 0, y_1, y_2).$$

Equations (11) and (12) can be written

$$(13) \quad \begin{cases} u_{x_1} + iu_{x_2} = 0 \\ \frac{\partial}{\partial x_2}(u_{x_1} + iu_{x_2}) = 0 \end{cases} \quad \text{at } (x_1, 0, y_1, y_2).$$

As before, we can also derive an equation for $u_{x_1} + iu_{x_2}$ and conclude that we must have $u_{x_1} + iu_{x_2} = 0$ everywhere. Similarly, we prove that $u_{y_1} + iu_{y_2} = 0$ everywhere. Hence u is complex analytic, and the real-valued solution of the original equation $u_{xx} + u_{yy} = f(x, y, u, p, q)$ is real analytic.

Now we are ready to tackle the general case. We consider an elliptic solution u of a general second order equation

$$0 = F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = F(x, y, u, p, q, r, s, t).$$

For convenience, we will often speak as if u were defined on all of \mathbb{R}^2 , although actually the arguments are entirely local. Ellipticity of u means that

$$0 < 4F_r F_t - F_s^2$$

[where F_r, F_s, F_t are evaluated at $(x, y, u(x, y), \dots, u_{yy}(x, y))$], so that, in particular, $F_r, F_t \neq 0$. Let ρ_1, ρ_2 be the two continuous everywhere unequal complex-valued functions which are solutions of

$$F_t - F_s \rho + F_r \rho^2 = 0,$$

and let

$$\begin{aligned} ' &= \frac{\partial}{\partial x} + \rho_1 \frac{\partial}{\partial y} \\ ^\wedge &= \frac{\partial}{\partial x} + \rho_2 \frac{\partial}{\partial y}. \end{aligned}$$

Then we automatically have

$$\begin{aligned} \rho_1 x' - y' &= 0 \\ \rho_2 x^\wedge - y^\wedge &= 0. \end{aligned}$$

Together with the considerations of the previous section we obtain eight equations

$$(*) \quad \begin{cases} \sum_{j=1}^8 a_{ij} \phi_j' = 0 & i = 1, \dots, 6 \\ \sum_{j=1}^8 a_{ij} \phi_j^\wedge = 0 & i = 7, 8, \end{cases}$$

where ϕ_1, \dots, ϕ_8 stand for x, y, u, \dots, t , and the a_{ij} are now complex-valued

functions of ϕ_1, \dots, ϕ_8 . Since F is assumed analytic in all its arguments, the functions a_{ij} are *complex analytic* in some region of \mathbb{C}^2 containing the set $\mathbb{R}^2 \subset \mathbb{C}^2$ where they are defined. We can arrange our equations as in the previous section, with the matrix (a_{ij}) being the matrix \mathcal{P} on page 95.

It will be convenient to use $\xi_1, \xi_2, \eta_1, \eta_2$ as coordinates on \mathbb{R}^4 . Thus we regard the ϕ_j as functions with $\phi_j(\xi_1, 0, \eta_1, 0)$ defined; in particular, we have

$$\begin{aligned} x(\xi_1, 0, \eta_1, 0) &= \phi_1(\xi_1, 0, \eta_1, 0) = \xi_1, \\ y(\xi_1, 0, \eta_1, 0) &= \phi_2(\xi_1, 0, \eta_1, 0) = \eta_1. \end{aligned}$$

The operators $'$ and \backslash in the (ξ_1, η_1) -plane = the (x, y) -plane are then given by

$$(*) \quad \begin{cases} ' = \frac{\partial}{\partial \xi_1} + \rho_1 \frac{\partial}{\partial \eta_1} \\ \backslash = \frac{\partial}{\partial \xi_1} + \rho_2 \frac{\partial}{\partial \eta_1}. \end{cases}$$

We consider the functions a_{ij} as already extended to complex analytic functions of their eight arguments in a suitable region of \mathbb{C}^2 . Now for fixed η_1 , consider equations $(*)$ as equations in the (ξ_1, η_2) -plane, with the operations $'$ and \backslash now being defined by

$$(*) \quad \begin{cases} ' = \frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \\ \backslash = -\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2}. \end{cases}$$

This is equivalent to taking $\rho_1 = 1$ and $\rho_2 = -1$. So we can write our equations as the matrix equation on page 95, with $\rho_1 = 1$ and $\rho_2 = -1$.

Setting

$$\begin{aligned} \phi_1 &= \psi_1 + i\psi_2 \\ \phi_2 &= \psi_3 + i\psi_4 & \psi_i \text{ real-valued,} \\ &\vdots \end{aligned}$$

and writing our equations in terms of the ψ_i , we obtain a matrix equation

$$\mathbf{P} \cdot \psi_{\eta_2} = \mathbf{Q} \cdot \psi_{\xi_1},$$

with

$$\mathbf{Q} = \mathbf{D}\mathbf{P},$$

where \mathbf{D} is obtained from \mathcal{D} simply by writing each row twice. So by Theorem 10, we can solve (*), with ' and \ given by (*₂); as our initial conditions we just choose

$$\phi_j(\xi_1, 0, \eta_1, 0) = \text{the original } \phi_j(\xi_1, 0, \eta_1, 0).$$

Similarly, we now extend the functions $\phi_j(\xi_1, 0, \eta_1, \eta_2)$ to \mathbb{R}^4 by fixing ξ_1 and η_2 , and considering equations (*) in the (ξ_2, η_1) -plane, with the operations ' and \ now defined by

$$(*)_3 \quad \begin{cases} ' = \frac{\partial}{\partial \xi_2} - \frac{\partial}{\partial \eta_1} \\ \backslash = \frac{\partial}{\partial \xi_2} + \frac{\partial}{\partial \eta_1}. \end{cases}$$

Among the extended functions ϕ_j , we have "x" = ϕ_1 and "y" = ϕ_2 . Since the ϕ_j are now complex-valued, we have four real-valued functions on \mathbb{R}^4 defined by

$$x = x_1 + ix_2, \quad y = y_1 + iy_2.$$

We claim that x_1, x_2, y_1, y_2 is a coordinate system in a neighborhood of any point in the (ξ_1, η_1) -plane. To prove this, we have to compute the Jacobian of (x_1, x_2, y_1, y_2) . First of all, since x_1, x_2, y_1, y_2 are simply $\xi_1, \xi_2, \eta_1, \eta_2$ on the (ξ_1, η_1) -plane, at the point in question we have

$$(1) \quad \begin{cases} \frac{\partial x_1}{\partial \xi_1} = 1, & \frac{\partial x_2}{\partial \xi_1} = 0, & \frac{\partial y_1}{\partial \xi_1} = 0, & \frac{\partial y_2}{\partial \xi_1} = 0 \\ \frac{\partial x_1}{\partial \eta_1} = 0, & \frac{\partial x_2}{\partial \eta_1} = 0, & \frac{\partial y_1}{\partial \eta_1} = 1, & \frac{\partial y_2}{\partial \eta_1} = 0. \end{cases}$$

To compute other derivatives, we first write the two complex-conjugate roots ρ_1, ρ_2 of $F_t - F_s \rho + F_r \rho^2 = 0$ as

$$\rho_1 = \sigma_1 + i\sigma_2, \quad \rho_2 = \sigma_1 - i\sigma_2, \quad \sigma_2 \neq 0.$$

The equations $y' - \rho_1 x' = 0$ and $y\backslash - \rho_2 x\backslash = 0$, with the two different meanings (*₂) and (*₃) for ' and \, give the following equations [after making use of (1)]:

$$(2) \quad \begin{cases} \frac{\partial y_1}{\partial \eta_2} + i \frac{\partial y_2}{\partial \eta_2} - (\sigma_1 + i\sigma_2) \left(1 + \frac{\partial x_1}{\partial \eta_2} + i \frac{\partial x_2}{\partial \eta_2} \right) = 0 \\ \frac{\partial y_1}{\partial \eta_2} + i \frac{\partial y_2}{\partial \eta_2} - (\sigma_1 - i\sigma_2) \left(-1 + \frac{\partial x_1}{\partial \eta_2} + i \frac{\partial x_2}{\partial \eta_2} \right) = 0 \end{cases}$$

$$(3) \quad \begin{cases} \frac{\partial y_1}{\partial \xi_2} + i \frac{\partial y_2}{\partial \xi_2} - 1 - (\sigma_1 + i\sigma_2) \left(\frac{\partial x_1}{\partial \xi_2} + i \frac{\partial x_2}{\partial \xi_2} \right) = 0 \\ \frac{\partial y_1}{\partial \xi_2} + i \frac{\partial y_2}{\partial \xi_2} + 1 - (\sigma_1 - i\sigma_2) \left(\frac{\partial x_1}{\partial \xi_2} + i \frac{\partial x_2}{\partial \xi_2} \right) = 0 \end{cases}$$

Subtracting the first equation of (2) from the second gives

$$\begin{aligned} 2\sigma_1 + 2i\sigma_2 \left(\frac{\partial x_1}{\partial \eta_2} + i \frac{\partial x_2}{\partial \eta_2} \right) = 0 &\implies \frac{\partial x_1}{\partial \eta_2} + i \frac{\partial x_2}{\partial \eta_2} = i \frac{\sigma_1}{\sigma_2} \\ &\implies \frac{\partial x_1}{\partial \eta_2} = 0, \quad \frac{\partial x_2}{\partial \eta_2} = \frac{\sigma_1}{\sigma_2}. \end{aligned}$$

Then we get

$$\begin{aligned} \frac{\partial y_1}{\partial \eta_2} + i \frac{\partial y_2}{\partial \eta_2} &= (\sigma_1 + i\sigma_2) \left(1 + i \frac{\sigma_1}{\sigma_2} \right) = \frac{i(\sigma_1^2 + \sigma_2^2)}{\sigma_2} \\ &\implies \frac{\partial y_1}{\partial \eta_2} = 0, \quad \frac{\partial y_2}{\partial \eta_2} = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2}. \end{aligned}$$

Similarly, from (3) we get

$$\begin{aligned} 2 + 2i\sigma_2 \left(\frac{\partial x_1}{\partial \xi_2} + i \frac{\partial x_2}{\partial \xi_2} \right) = 0 &\implies \frac{\partial x_1}{\partial \xi_2} + i \frac{\partial x_2}{\partial \xi_2} = \frac{i}{\sigma_2} \\ &\implies \frac{\partial x_1}{\partial \xi_2} = 0, \quad \frac{\partial x_2}{\partial \xi_2} = \frac{1}{\sigma_2}, \end{aligned}$$

and then

$$\begin{aligned} \frac{\partial y_1}{\partial \xi_2} + i \frac{\partial y_2}{\partial \xi_2} &= 1 + (\sigma_1 + i\sigma_2) \frac{i}{\sigma_2} = i \frac{\sigma_1}{\sigma_2} \\ &\implies \frac{\partial y_1}{\partial \xi_2} = 0, \quad \frac{\partial y_2}{\partial \xi_2} = \frac{\sigma_1}{\sigma_2}. \end{aligned}$$

So at the point in question, the matrix of derivatives of x_1, x_2, y_1, y_2 with respect to $\xi_1, \xi_2, \eta_1, \eta_2$ is

$$(4) \quad \begin{array}{c} \begin{matrix} x_1 & x_2 & y_1 & y_2 \end{matrix} \\ \begin{matrix} \xi_1 \\ \xi_2 \\ \eta_1 \\ \eta_2 \end{matrix} \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{1}{\sigma_2} & 0 & \frac{\sigma_1}{\sigma_2} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{\sigma_1}{\sigma_2} & 0 & \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2} \end{pmatrix}.$$

The determinant equals 1, so (x_1, x_2, y_1, y_2) is indeed a coordinate system.

Now all partials $\partial/\partial \xi_1, \dots, \partial/\partial \eta_2$ can be written as certain linear combinations of $\partial/\partial x_1, \dots, \partial/\partial y_2$. So we can also write

$$(5) \quad \frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} = A_1 \frac{\partial}{\partial x_1} + B_1 \frac{\partial}{\partial y_1} + C_1 \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + D_1 \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right).$$

Now the equation $y' - \rho_1 x' = 0$, where $'$ has the significance $(*_2)$, gives

$$(6) \quad \left(\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) (y_1 + iy_2) - \rho_1 \left(\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) (x_1 + ix_2) = 0.$$

But $\partial/\partial x_1 + i\partial/\partial x_2$ gives zero when applied to $x_1 + ix_2$ (or any analytic function of x_1, x_2), and similarly for $\partial/\partial y_1 + i\partial/\partial y_2$. So when we replace the operator $\partial/\partial \xi_1 + \partial/\partial \eta_2$ in equation (6) by its expression in (5) we end up with

$$B_1 - \rho_1 A_1 = 0.$$

Note that if we had $A_1 = 0$, then the operator (5) would not be real unless $C_1 = D_1 = 0$, which is impossible, since $\partial/\partial \xi_1 \neq -\partial/\partial \eta_2$. So $A_1 \neq 0$. Thus we have

$$(7) \quad \frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} = A_1 \frac{\partial}{\partial x_1} + \rho_1 A_1 \frac{\partial}{\partial y_1} + C_1 \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + D_1 \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right),$$

and similarly

$$(8) \quad -\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} = A_2 \frac{\partial}{\partial x_1} + \rho_2 A_2 \frac{\partial}{\partial y_1} + C_2 \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + D_2 \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right),$$

$$(9) \quad \frac{\partial}{\partial \xi_2} - \frac{\partial}{\partial \eta_1} = E_1 \frac{\partial}{\partial x_1} + \rho_1 E_1 \frac{\partial}{\partial y_1} + G_1 \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + H_1 \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right),$$

$$(10) \quad \frac{\partial}{\partial \xi_2} + \frac{\partial}{\partial \eta_1} = E_2 \frac{\partial}{\partial x_1} + \rho_2 E_2 \frac{\partial}{\partial y_1} + G_2 \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + H_2 \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right),$$

where $A_1, A_2, E_1, E_2 \neq 0$. All quantities A_1, \dots, H_2 are simply linear combinations of the derivatives of x_1, x_2, y_1, y_2 with respect to $\xi_1, \xi_2, \eta_1, \eta_2$. For

example, we obviously have

$$\begin{aligned}
 iC_1 &= \left(\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) x_2, & iD_1 &= \left(\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) y_2, \\
 iC_2 &= \left(-\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) x_2, & iD_2 &= \left(-\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \right) y_2, \\
 iG_1 &= \left(\frac{\partial}{\partial \xi_2} - \frac{\partial}{\partial \eta_1} \right) x_2, & iH_1 &= \left(\frac{\partial}{\partial \xi_2} - \frac{\partial}{\partial \eta_1} \right) y_2, \\
 iG_2 &= \left(\frac{\partial}{\partial \xi_2} + \frac{\partial}{\partial \eta_1} \right) x_2, & iH_2 &= \left(\frac{\partial}{\partial \xi_2} + \frac{\partial}{\partial \eta_1} \right) y_2.
 \end{aligned}$$

In particular, at a point in the (ξ_1, η_1) -plane we have, from the entries of the matrix (4),

$$\begin{aligned}
 iC_1 &= \frac{\sigma_1}{\sigma_2}, & iD_1 &= \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2}, \\
 iC_2 &= \frac{\sigma_1}{\sigma_2}, & iD_2 &= \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2}, \\
 iG_1 &= \frac{1}{\sigma_2}, & iH_1 &= \frac{\sigma_1}{\sigma_2}, \\
 iG_2 &= \frac{1}{\sigma_2}, & iH_2 &= \frac{\sigma_1}{\sigma_2}.
 \end{aligned}
 \tag{11}$$

Notice that up till now we have used only the two simplest equations of (*). We will now use the whole set. In the initial plane, the equations (*) hold in three different forms, corresponding to the three meanings of the operators ' and ', namely

$$\begin{aligned}
 (*) \quad \left\{ \begin{array}{l} ' = \frac{\partial}{\partial x_1} + \rho_1 \frac{\partial}{\partial y_1} \\ ' = \frac{\partial}{\partial x_1} + \rho_2 \frac{\partial}{\partial y_1} \end{array} \right. & \quad (*) \quad \left\{ \begin{array}{l} ' = \frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \\ ' = -\frac{\partial}{\partial \xi_1} + \frac{\partial}{\partial \eta_2} \end{array} \right. & \quad (*) \quad \left\{ \begin{array}{l} ' = \frac{\partial}{\partial \xi_2} - \frac{\partial}{\partial \eta_1} \\ ' = \frac{\partial}{\partial \xi_2} + \frac{\partial}{\partial \eta_1} \end{array} \right.
 \end{aligned}$$

From the equations with (*) we have, making use of (7) and (11),

$$\begin{aligned}
 (12) \quad \sum_j a_{ij} \left[A_1 \frac{\partial}{\partial x_1} + \rho_1 A_1 \frac{\partial}{\partial y_1} \right. \\
 \left. + \frac{\sigma_1}{i\sigma_2} \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + \frac{\sigma_1^2 + \sigma_2^2}{i\sigma_2} \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right) \right] \phi_j = 0
 \end{aligned}$$

for $i = 1, \dots, 6$. From the equations with $(*_1)$, we have, after multiplying by A_1 ,

$$(13) \quad \sum_j a_{ij} \left[A_1 \frac{\partial}{\partial x_1} + \rho_1 A_1 \frac{\partial}{\partial y_1} \right] \phi_j = 0$$

for $i = 1, \dots, 6$. Subtracting (13) from (12) gives

$$(14) \quad \sum_j a_{ij} \left[\frac{\sigma_1}{i\sigma_2} \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + \frac{\sigma_1^2 + \sigma_2^2}{i\sigma_2} \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right) \right] \phi_j = 0.$$

If we do the same thing for $i = 7, 8$, except multiply by A_2 instead of A_1 , we find that (14) holds also for $i = 7, 8$. Since $\det(a_{ij}) \neq 0$, it follows that

$$(15) \quad \left[\frac{\sigma_1}{\sigma_2} \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2} \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right) \right] \phi_j = 0 \quad j = 1, \dots, 8.$$

Similarly, if we start from the equations with $(*_3)$, and then subtract the equations with $(*_1)$, multiplied by E_1 and E_2 , we find that

$$(16) \quad \left[\frac{1}{\sigma_2} \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) + \frac{\sigma_1}{\sigma_2} \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right) \right] \phi_j = 0 \quad j = 1, \dots, 8.$$

For each particular ϕ_j , equations (15) and (16) give two equations for ϕ_j , and since

$$\det \begin{pmatrix} \frac{\sigma_1}{\sigma_2} & \frac{\sigma_1^2 + \sigma_2^2}{\sigma_2} \\ \frac{1}{\sigma_2} & \frac{\sigma_1}{\sigma_2} \end{pmatrix} = -1 \neq 0,$$

we must have

$$(17) \quad \left(\frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right) \phi_j = 0 \quad \text{and} \quad \left(\frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2} \right) \phi_j = 0.$$

Thus, we see that the Cauchy-Riemann equations for ϕ_j hold in the plane $x_2 = y_2 = 0$.

Now we want to show that the Cauchy-Riemann equations hold for the y_1, y_2 variables. Let

$$\nabla_x = \frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2}, \quad \nabla_y = \frac{\partial}{\partial y_1} + i \frac{\partial}{\partial y_2}.$$

We denote the partials of the functions a_{ij} with respect to their 8 variables by $\partial a_{ij} / \partial \phi_l$, etc. Because the a_{ij} are *analytic*, we have

$$\nabla_x a_{ij} = \sum_l \frac{\partial a_{ij}}{\partial \phi_l} \nabla_x \phi_l, \quad \nabla_y a_{ij} = \sum_l \frac{\partial a_{ij}}{\partial \phi_l} \nabla_y \phi_l.$$

Consider the first 6 equations (*), with ' and ` given by (*₂); after division by A_1 they can be written

$$\sum_j a_{ij} \frac{\phi_j'}{A_1} = \sum_j a_{ij} \left[\frac{\partial}{\partial x_1} + \rho_1 \frac{\partial}{\partial y_1} + \frac{C_1}{A_1} \nabla_x + \frac{D_1}{A_1} \nabla_y \right] \phi_j = 0 \quad i = 1, \dots, 6.$$

Apply ∇_x to this equation. Since we have

$$\begin{aligned} \nabla_x \frac{\phi_j'}{A_1} &= \nabla_x \left(\left(\frac{\partial}{\partial x_1} + \rho_1 \frac{\partial}{\partial y_1} + \frac{C_1}{A_1} \nabla_x + \frac{D_1}{A_1} \nabla_y \right) \phi_j \right) \\ &= \left(\frac{\partial}{\partial x_1} + \rho_1 \frac{\partial}{\partial y_1} + \frac{C_1}{A_1} \nabla_x + \frac{D_1}{A_1} \nabla_y \right) \nabla_x \phi_j \\ &\quad + \left(\nabla_x \frac{C_1}{A_1} \right) \cdot \nabla_x \phi_j + \left(\nabla_x \frac{D_1}{A_1} \right) \cdot \nabla_y \phi_j + \frac{\partial \phi_j}{\partial y_1} \sum_{l=1}^8 \frac{\partial \rho_1}{\partial \phi_l} \nabla_x \phi_l \\ &= \frac{1}{A_1} (\nabla_x \phi_j)' + \left(\nabla_x \frac{C_1}{A_1} \right) \cdot \nabla_x \phi_j + \left(\nabla_x \frac{D_1}{A_1} \right) \cdot \nabla_y \phi_j + \frac{\partial \phi_j}{\partial y_1} \sum_{l=1}^8 \frac{\partial \rho_1}{\partial \phi_l} \nabla_x \phi_l, \end{aligned}$$

we obtain an equation of the form

$$(18) \quad \sum_j a_{ij} (\nabla_x \phi_j)' + \sum_j (b_{ij} \nabla_x \phi_j + c_{ij} \nabla_y \phi_j) = 0 \quad i = 1, \dots, 6.$$

Treating the equations for $i = 7, 8$ similarly, except dividing by A_2 , we obtain

$$(19) \quad \sum_j a_{ij} (\nabla_x \phi_j)' + \sum_j (b_{ij} \nabla_x \phi_j + c_{ij} \nabla_y \phi_j) = 0 \quad i = 7, 8.$$

Applying ∇_y similarly to these same equations, we obtain

$$(20) \quad \sum_j a_{ij} (\nabla_y \phi_j)' + \sum_j (d_{ij} \nabla_x \phi_j + e_{ij} \nabla_y \phi_j) = 0 \quad i = 1, \dots, 6$$

$$(21) \quad \sum_j a_{ij} (\nabla_y \phi_j)' + \sum_j (d_{ij} \nabla_x \phi_j + e_{ij} \nabla_y \phi_j) = 0 \quad i = 7, 8.$$

Equations (18)–(21) are 16 equations for 16 complex-valued functions $\nabla_x \phi_j$, $\nabla_y \phi_j$. The matrix of the system is

$$\begin{pmatrix} (a_{ik}) & 0 \\ 0 & (a_{ik}) \end{pmatrix}.$$

So we easily see that the corresponding system of 32 equations for 32 real-valued functions is hyperbolic. But we know from (17) that $\nabla_x \phi_j = \nabla_y \phi_j = 0$ for $\eta_2 = 0$. By uniqueness of solutions, it follows that $\nabla_x \phi_j = \nabla_y \phi_j$ for all $(\xi_1, 0, \eta_1, \eta_2)$.

In exactly the same way, we show finally that $\nabla_x \phi_j = \nabla_y \phi_j = 0$ for all $(\xi_1, \xi_2, \eta_1, \eta_2)$. Thus all extended ϕ_j , in particular $u = \phi_3$, are complex analytic. So the original real solution u of our equation is real analytic.

In this proof we need the ϕ_j to have continuous second partial derivatives satisfying a Lipschitz condition (so that the $\nabla_x \phi_j, \nabla_y \phi_j$ in the last step will have continuous partials satisfying a Lipschitz condition). Thus we require u to have continuous fourth partial derivatives satisfying a Lipschitz condition. Actually, the result holds even if u is C^3 , but that information comes out of other proofs (it might also be derivable from the present proof with enough extra work). We will merely state this stronger result in the summary of all the work of this section:

13. THEOREM. If u is a C^3 elliptic solution of the equation

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0,$$

where F is a real analytic function of its 8 arguments, then u is real analytic.

ADDENDUM 1

DIFFERENTIAL SYSTEMS;
THE CARTAN-KÄHLER THEOREM

Suppose we are given everywhere linearly independent 1-forms $\omega_1, \dots, \omega_l$ on an n -manifold M . The Frobenius integrability theorem, in the differential form version (Proposition I.7-14), tells us when every point $p \in M$ lies in some $(n-l)$ -dimensional manifold $N \subset M$ such that all ω_j restricted to N are zero: this happens if and only if each $d\omega_j$ is in the ideal generated by the $\{\omega_j\}$. Our proof rested on the observation that the $d\omega_j$ have this property if and only if the $(n-l)$ -dimensional distribution $\Delta = \bigcap_k \ker \omega_j$ has the property that $[X, Y]$ belongs to Δ whenever X and Y do. On the other hand, simple direct considerations could have shown us that the condition on the $d\omega_j$ is certainly necessary. For suppose that $N \subset M$ is an $(n-l)$ -dimensional submanifold of M on which all ω_j vanish (i.e., $i^*\omega_j = 0$, where $i: N \rightarrow M$ is the inclusion map). Then the $d\omega_j$ also vanish on N , since

$$i^*(d\omega_j) = d(i^*\omega_j) = 0.$$

But the 2-forms $\omega_j \wedge \omega_k$ also vanish on N , and because the ω_j are everywhere linearly independent, at each point $p \in N$ the $\{\omega_j(p) \wedge \omega_k(p)\}$ already span the set of all elements of $\Omega^2(M_p)$ which vanish on N_p . Thus $d\omega_j(p)$ must be a linear combination of the $\{\omega_j(p) \wedge \omega_k(p)\}$.

We could also have given a direct proof that this necessary condition is sufficient, without appealing to the first version of the Frobenius integrability theorem. We will briefly outline this proof, for it not only shows just how the condition on the $d\omega_i$ is related to the classical integrability criterion, but it is also similar in approach to the proof of the main theorem which we will be proving later.

For convenience we set $k = n - l$, and number our forms as $\omega_{k+1}, \dots, \omega_n$. Since the result is essentially local, we can assume that $M = \mathbb{R}^n$, that the point $p \in M$ in question is $0 \in \mathbb{R}^n$, and, by changing our axes if necessary, that $dx^1, \dots, dx^k, \omega_{k+1}, \dots, \omega_n$ span $(\mathbb{R}_0^n)^*$. This means that near 0 we can write

$$(1) \quad dx^\rho = \sum_{h=1}^k A_{h\rho} dx^h + \sum_{r=k+1}^n B_{r\rho} \omega_r \quad \rho = k+1, \dots, n.$$

Now take d of equation (1), and consider the coefficient of a term $dx^i \wedge dx^j$ ($i < j \leq k$), when the right side is expressed in terms of the 2-forms

$$dx^i \wedge dx^j, \quad dx^i \wedge \omega_r, \quad \omega_r \wedge \omega_s,$$

which are linearly independent near 0. When we write $d\omega_r$ in this way, the coefficients of $dx^i \wedge dx^j$ must vanish, since by hypothesis $d\omega_r$ is in the ideal generated by the ω_r . So we obtain

$$\begin{aligned}
 0 &= \text{coefficient of } dx^i \wedge dx^j \text{ in } \sum_{h=1}^k dA_{h\rho} \wedge dx^h \\
 &= \text{coefficient of } dx^i \wedge dx^j \text{ in } \sum_{h=1}^k \sum_{\sigma=1}^n \frac{\partial A_{h\rho}}{\partial x^\sigma} dx^\sigma \wedge dx^h \\
 &= \frac{\partial A_{j\rho}}{\partial x^i} - \frac{\partial A_{i\rho}}{\partial x^j} \\
 &\quad + \text{coefficient of } dx^i \wedge dx^j \text{ in } \sum_{h=1}^k \sum_{\sigma=k+1}^n \sum_{t=1}^k \frac{\partial A_{h\rho}}{\partial x^\sigma} A_{t\sigma} dx^t \wedge dx^h \\
 &\quad \text{by (1)}
 \end{aligned}$$

and thus, finally,

$$(2) \quad 0 = \frac{\partial A_{j\rho}}{\partial x^i} - \frac{\partial A_{i\rho}}{\partial x^j} + \sum_{\sigma=k+1}^n \frac{\partial A_{j\rho}}{\partial x^\sigma} A_{i\sigma} - \sum_{\sigma=k+1}^n \frac{\partial A_{i\rho}}{\partial x^\sigma} A_{j\sigma}.$$

But now the classical integrability result (Theorem I.6-1) shows that we can find functions f^{k+1}, \dots, f^n in a neighborhood of 0 in \mathbb{R}^k such that

$$(3) \quad \frac{\partial f^\rho}{\partial x^h}(x_1, \dots, x_k) = A_{h\rho}(x_1, \dots, x_k, f^{k+1}(x_1, \dots, x_k), \dots, f^n(x_1, \dots, x_k)).$$

Equation (3) is precisely the condition that the ω_r vanish on the submanifold $\{(x_1, \dots, x_k, f^{k+1}(x_1, \dots, x_k), \dots, f^n(x_1, \dots, x_k))\}$, so the proof is complete.

Now we want to consider a more general question. Suppose we are given an ideal \mathfrak{d} of differential forms on M , not necessarily generated by 1-forms, which satisfies $d\mathfrak{d} \subset \mathfrak{d}$. When is there a submanifold $N \subset M$ such that all forms of \mathfrak{d} vanish on N ? We warn right away that everything is going to be much more complicated. The basic information regarding this situation is contained in the Cartan-Kähler theorem (first proved by Cartan when \mathfrak{d} is generated by 1-forms and 2-forms, and then generalized by Kähler). We will never use this result, except to give an alternative proof of a theorem, in the Addendum to Chapter 11, but I felt that it should be included here, not only because it is an application of the Cauchy-Kowalewski theorem, but also because it plays such a crucial role in the work of É. Cartan. It enables one to say, in a sense that will

be clarified later on, “how many” different submanifolds of \mathbb{R}^n satisfy a given geometric condition, e.g., the condition that H is constant [here we are considering the *local* theory of submanifolds, without any completeness requirements]; numerous such examples are worked out in É. Cartan {2}. The Cartan-Kähler theorem may be thought of as a result about integrability conditions for systems of partial differential equations, of a more complex type than (3). Nevertheless, the systems to be considered are still very special, since they come from differential forms—one could compare this situation with the Poincaré Lemma, which also involves integrability conditions of a very special sort.

Before we can state the Cartan-Kähler theorem, some preliminary definitions will be required. First we want to be more precise about ideals of differential forms. Let $\Omega^k(M)$ be the vector space of all k -forms on M . Then the direct sum $\Omega(M) = \Omega^0(M) \oplus \cdots \oplus \Omega^n(M)$ is a ring under \wedge . For any ideal $\mathcal{I} \subset \Omega(M)$, we set $\mathcal{I}_k = \mathcal{I} \cap \Omega^k(M)$. We will consider only ideals \mathcal{I} which are **homogeneous**, meaning that

$$\mathcal{I} = \mathcal{I}_0 \oplus \mathcal{I}_1 \oplus \cdots \oplus \mathcal{I}_n.$$

Thus, for example, if \mathcal{I} contains $\omega_1 + \omega_2$ where ω_1 is a 1-form and ω_2 is a 2-form, then \mathcal{I} must contain ω_1 and ω_2 (so \mathcal{I} could not be the ideal generated by $\omega_1 + \omega_2$). For a homogeneous ideal \mathcal{I} it is certainly clear what we mean by the condition $d\mathcal{I} \subset \mathcal{I}$: for each k -form $\omega \in \mathcal{I}$, the $(k+1)$ -form $d\omega$ must also be in \mathcal{I} . A homogeneous ideal with this property is called a **differential ideal**, or sometimes a **differential system**. For the present we will assume that our differential system \mathcal{I} does not contain functions, i.e., that $\mathcal{I}_0 = 0$.

Let \mathcal{I} be any homogeneous ideal with $\mathcal{I}_0 = 0$ (not necessarily satisfying $d\mathcal{I} \subset \mathcal{I}$). An l -dimensional submanifold $N \subset M$, with inclusion map $i: N \rightarrow M$, is called an **integral submanifold** of \mathcal{I} if $i^*\omega = 0$ for all forms $\omega \in \mathcal{I}$. It is easy to see that, because \mathcal{I} is an ideal, this condition holds if $i^*\omega = 0$ for all forms $\omega \in \mathcal{I}_l$. It is also easy to see that if \mathcal{I} is generated by a set of elements S , then it suffices to have $i^*\omega = 0$ for all $\omega \in S$ of degree $\leq l$. In order to analyze integral submanifolds of \mathcal{I} , we consider the possible tangent spaces for them. An l -dimensional subspace $W \subset M_p$ of M_p is called an (**l -dimensional**) **integral element** of \mathcal{I} if all $\omega(p)$ are zero when restricted to W , for all $\omega \in \mathcal{I}$; again, it suffices to have this for all $\omega \in \mathcal{I}_l$, or for all ω of degree $\leq l$ in a generating set S . Notice that a subspace of an integral element is also an integral element. We will also allow the 0-dimensional subspace of M_p , which we will identify with p . It is always an integral element, since we are assuming that $\mathcal{I}_0 = 0$.

When the ideal \mathcal{I} is generated by 1-forms, we must assume, for the Frobenius integrability theorem, that locally \mathcal{I} is generated by a fixed number of linearly independent 1-forms. The analogous requirements for an arbitrary differential

system \mathcal{I} are more involved. Let $W \subset M_p$ be a k -dimensional integral element, and let X_1, \dots, X_k be any basis. We define the “polar space”

$$\mathcal{E}(W) = \{X \in M_p : \omega(p)(X_1, \dots, X_k, X) = 0 \text{ for all } \omega \in \mathcal{I}_{k+1}\}.$$

[For $k = 0$, this means that $\mathcal{E}(p) = \{X \in M_p : \omega(p)(X) = 0 \text{ for all } \omega \in \mathcal{I}_1\}$.] This definition is clearly independent of the basis X_1, \dots, X_k , and we have $W \subset \mathcal{E}(W)$. Using the fact that \mathcal{I} is an *ideal*, we easily see that for all $X \in \mathcal{E}(W)$ and all $h \leq k$ we have

$$\omega(p)(X_{i_1}, \dots, X_{i_h}, X) = 0 \quad \text{for all } \omega \in \mathcal{I}_{h+1}.$$

This means that for every $X \in \mathcal{E}(W)$ which is not in W , the space $W \oplus \mathbb{R} \cdot X$ is an extension of W to a $(k + 1)$ -dimensional integral element; conversely, any $(k + 1)$ -dimensional integral element extending W is of this form. We will also find it useful to consider explicitly the ordered bases (X_1, \dots, X_k) of integral elements. Let

$$\begin{aligned} \mathcal{M}_k &= \{(p, X_1, \dots, X_k) : X_1, \dots, X_k \text{ span a} \\ &\quad k\text{-dimensional integral element of } M_p\} \\ &\subset M \times TM \times \dots \times TM. \end{aligned}$$

For each $(p, X_1, \dots, X_k) \in \mathcal{M}_k$, we define

$$\begin{aligned} \mathcal{E}(p, X_1, \dots, X_k) &= \{X \in M_p : \omega(p)(X_1, \dots, X_k, X) = 0 \text{ for all } \omega \in \mathcal{I}_{k+1}\} \\ &= \mathcal{E}(k\text{-dimensional integral element spanned by } X_1, \dots, X_k). \end{aligned}$$

We now define **regular** integral elements inductively as follows. The point p is a **regular** 0-dimensional integral element if $\dim \mathcal{E}_1(p') = \dim \mathcal{E}_1(p)$ for all p' in a neighborhood of p . A k -dimensional integral element W is **regular** if

- (a) W contains a $(k - 1)$ -dimensional regular integral element,
- (b) $\dim \mathcal{E}_{k+1}(W') = \dim \mathcal{E}_{k+1}(W)$ for all k -dimensional integral elements W' in a neighborhood of W .

In order to talk about a neighborhood of W , we have to specify the topology involved. The k -dimensional integral elements are topologized as a subset of the set of all k -dimensional subspaces of all M_q ; locally this looks like $\mathbb{R}^n \times (k\text{-dimensional subspaces of } \mathbb{R}^n)$, and we use the obvious topology on k -dimensional subspaces of \mathbb{R}^n (described in detail in Chapter 13, section 2).

Equivalently, W is regular if it has some basis X_1, \dots, X_k such that for each $h \leq k$ we have $\dim \mathcal{E}_{h+1}(p', X'_1, \dots, X'_h) = \dim \mathcal{E}_{h+1}(p, X_1, \dots, X_h)$ for all $(p', X'_1, \dots, X'_h) \in \mathcal{M}_h$ in a neighborhood of (p, X_1, \dots, X_h) . Notice that the definition does not preclude the possibility that the regular k -dimensional integral element W contains a $(k-1)$ -dimensional integral element which is not regular. That is why our second criterion for regularity merely requires the existence of *some* basis (X_1, \dots, X_k) with the requisite property—there may also be bases which do not have this property. A basis (X_1, \dots, X_k) which does have the required property will be called **good**.

For a k -dimensional integral element W , consider the codimension

$$c_{k+1}(W) = n - \dim \mathcal{E}_{k+1}(W);$$

similarly, for $(p, X_1, \dots, X_k) \in \mathcal{M}_k$, set

$$c_{k+1}(p, X_1, \dots, X_k) = n - \dim \mathcal{E}_{k+1}(p, X_1, \dots, X_k).$$

Clearly $c_{k+1}(W)$ is the maximum number of $(k+1)$ -forms $\omega^{(1)}, \omega^{(2)}, \dots \in \mathcal{L}_{k+1}$ such that the $c_{k+1}(W)$ linear functions

$$Y \mapsto \omega^{(\alpha)}(p)(X_1, \dots, X_k, Y) \quad Y \in M_p \quad ((X_1, \dots, X_k) \text{ a basis of } W)$$

are linearly independent. It follows that the function $W \mapsto c_{k+1}(W)$ is lower semi-continuous on the set of all k -dimensional integral elements [that is, the value of this function may be greater than $c_{k+1}(W)$ arbitrarily close to W , but it cannot be less than $c_{k+1}(W)$ arbitrarily close to W]. Consequently, the function $W \mapsto \dim \mathcal{E}_{k+1}(W)$ is *upper* semi-continuous. It follows easily that condition (b) holds on an open dense subset of the set of all k -dimensional integral elements. It certainly holds if $\dim \mathcal{E}_{k+1}(W)$ has the minimum possible value. [In particular, condition (b) holds if $\dim \mathcal{E}_{k+1}(W) = 0$, in which case there is no $(k+1)$ -dimensional integral element containing W .] It is easy to see that if M is a connected analytic manifold, and we consider only *analytic* forms, then condition (b) is equivalent to $\dim \mathcal{E}_{k+1}(W)$ having the minimum possible value.

The appropriateness of the regularity condition is attested to by the following

14. LEMMA. Let \mathcal{L} be a homogeneous ideal with $\mathcal{L}_0 = 0$. If $X_1, \dots, X_k \in M_p$ is a good basis for a regular k -dimensional integral element of \mathcal{L} , and $X_{k+1} \in \mathcal{E}_{k+1}(p, X_1, \dots, X_k)$ is linearly independent of X_1, \dots, X_k , then near (p, X_1, \dots, X_{k+1}) , the set \mathcal{M}_{k+1} is a submanifold of $M \times TM \times \dots \times TM$, of dimension

$$n(k+2) - c_1(p) - c_2(p, X_1) - \dots - c_{k+1}(p, X_1, \dots, X_k).$$

PROOF. We can assume that $M = \mathbb{R}^n$. Recall that for $Y \in \mathbb{R}^n$, we let $Y_q = (q, Y)$ be the corresponding tangent vector $\in \mathbb{R}^n_q$. Choose A_1, \dots, A_k with $X_i = (A_i)_p$. Set

$$\tilde{\mathcal{M}}_{k+1} = \{(q, Y_1, \dots, Y_{k+1}) \in \mathbb{R}^{n(k+2)} : (q, (Y_1)_q, \dots, (Y_{k+1})_q) \in \mathcal{M}_{k+1}\}.$$

Then \mathcal{M}_{k+1} is the image of $\tilde{\mathcal{M}}_{k+1}$ under an imbedding $\mathbb{R}^{n(k+2)} \rightarrow \mathbb{R}^n \times T\mathbb{R}^n \times \dots \times T\mathbb{R}^n$, so it suffices to prove that $\tilde{\mathcal{M}}_{k+1}$ is a manifold. We will use induction on k , the case $k = 0$ being easy. So suppose that $\tilde{\mathcal{M}}_h \subset \mathbb{R}^{n(h+1)}$ is known to be a submanifold, of dimension

$$(1) \quad \dim \tilde{\mathcal{M}}_h = n(h+1) - c_1(p) - \dots - c_h(p, X_1, \dots, X_{h-1}).$$

For convenience, set

$$c_{h+1} = c_{h+1}(p, X_1, \dots, X_h).$$

Choose c_{h+1} $(h+1)$ -forms $\omega^{(1)}, \omega^{(2)}, \dots \in \mathfrak{L}_{h+1}$ such that the c_{h+1} linear functions

$$(*) \quad Y \mapsto \omega^{(\alpha)}(p)(X_1, \dots, X_h, Y_p)$$

are linearly independent. We adopt the convention that if $q, Y_1, \dots, Y_{h+1} \in \mathbb{R}^n$, then

$$\omega^{(\alpha)}(q, Y_1, \dots, Y_{h+1}) \quad \text{denotes} \quad \omega^{(\alpha)}(q)((Y_1)_q, \dots, (Y_{h+1})_q).$$

Thus we can consider $\omega^{(\alpha)}$ as a function on $\mathbb{R}^{n(h+2)}$. Since X_1, \dots, X_h is a good basis, we know that for $(q, Y_1, \dots, Y_h) \in \tilde{\mathcal{M}}_h$ close to (p, A_1, \dots, A_h) , the linear functions

$$Y \mapsto \omega^{(\alpha)}(q, Y_1, \dots, Y_h, Y)$$

already span the set of linear functions

$$Y \mapsto \omega(q, Y_1, \dots, Y_h, Y) \quad \text{for all } \omega \in \mathfrak{L}_{h+1}.$$

This means that near (p, A_1, \dots, A_h) , the set $\tilde{\mathcal{M}}_{h+1}$ is precisely the set of $(q, Y_1, \dots, Y_h, Y_{h+1})$ such that

$$\begin{cases} (q, Y_1, Y_h) \in \tilde{\mathcal{M}}_h \\ \omega^{(\alpha)}(q, Y_1, \dots, Y_h, Y_{h+1}) = 0 \end{cases} \quad \alpha = 1, \dots, c_{h+1}.$$

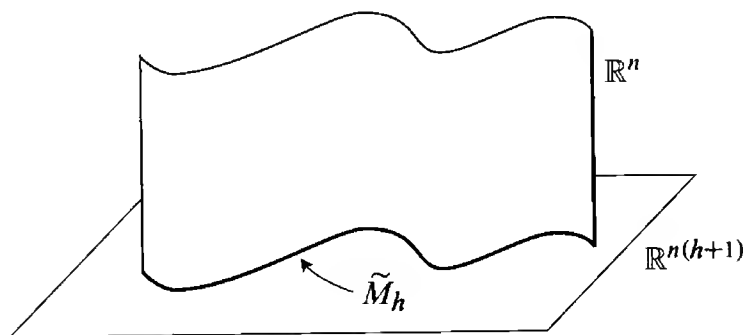
Thus, if we define

$$F: \tilde{\mathcal{M}}_h \times \mathbb{R}^n \rightarrow \mathbb{R}^{c_{h+1}}$$

by

$$\begin{aligned} F(q, Y_1, \dots, Y_h, Y_{h+1}) \\ = (\omega^{(1)}(q, Y_1, \dots, Y_h, Y_{h+1}), \omega^{(2)}(q, Y_1, \dots, Y_h, Y_{h+1}), \dots), \end{aligned}$$

then $\tilde{\mathcal{M}}_{h+1}$ is just $F^{-1}(0)$ near $(p, A_1, \dots, A_h, A_{h+1})$. Let Z_1, \dots, Z_n denote the



last n basis vectors of $\mathbb{R}^{n(h+2)}$. Then the linear independence of the functions (*) shows that the vectors

$$F_*((Z_i)_{(p, A_1, \dots, A_h, A_{h+1})}) \in (\mathbb{R}^{c_{h+1}})_0$$

are linearly independent. Thus F_* has rank c_{h+1} at $(p, A_1, \dots, A_h, A_{h+1})$. So in a neighborhood of $(p, A_1, \dots, A_h, A_{h+1})$ the set

$$\tilde{\mathcal{M}}_{h+1} = F^{-1}(0) \subset \tilde{\mathcal{M}}_h \times \mathbb{R}^n$$

is a manifold, of dimension

$$\begin{aligned} \dim \tilde{\mathcal{M}}_{h+1} &= \dim(\tilde{\mathcal{M}}_h \times \mathbb{R}^n) - c_{h+1} \\ &= n(h+2) - c_1(p) - \dots - c_h(p, X_1, \dots, X_{h-1}) - c_{h+1}, \quad \text{by (I). } \spadesuit \end{aligned}$$

Our goal is to show that if our ideal \mathcal{I} is a differential system ($d\mathcal{I} \subset \mathcal{I}$), then, at least in the analytic case, a k -dimensional integral element at p which contains a $(k-1)$ -dimensional regular integral element (but which need not be regular itself), is the tangent space at p of some k -dimensional integral submanifold of \mathcal{I} . We will derive this result as a corollary of a more precise one, which tells when a k -dimensional integral submanifold of \mathcal{I} can be extended to a $(k+1)$ -dimensional integral submanifold.

Suppose $W \subset M_p$ is a k -dimensional regular integral element of \mathcal{I} which is the tangent space at p of some k -dimensional integral submanifold N of \mathcal{I} .

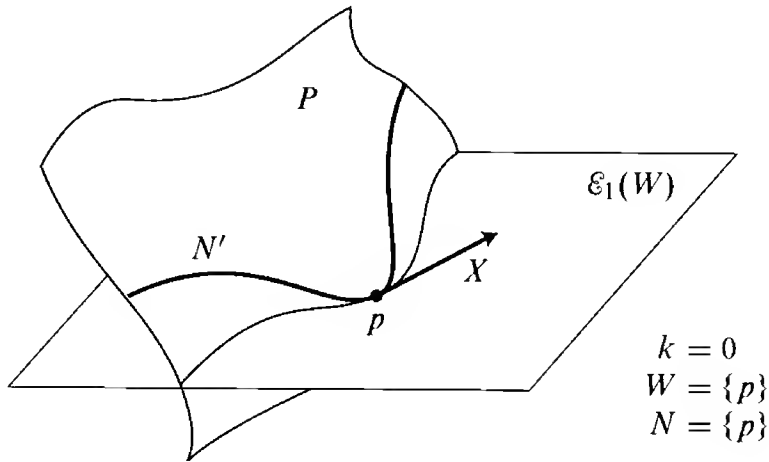
Suppose that $\dim \mathcal{E}_{k+1}(W) > k$, so that there is a vector $X \in \mathcal{E}_{k+1}(W)$ which is not in W ; then $W \oplus \mathbb{R} \cdot X$ is a $(k+1)$ -dimensional integral element. We will show that there is a $(k+1)$ -dimensional integral manifold $N' \supset N$ whose tangent space at p is $W \oplus \mathbb{R} \cdot X$. We can also say precisely how many such integral manifolds N' there are. To do this, we choose a submanifold P of M of dimension

$$\dim P = k + 1 + c_{k+1}(W),$$

such that

- (a) $P \supset N$
- (b) $P_p \cap \mathcal{E}_{k+1}(W) = W \oplus \mathbb{R} \cdot X$.

We will show that near p there is an (essentially unique) $(k+1)$ -dimensional integral submanifold N' of \mathcal{I} with $N \subset N' \subset P$ and $N'_p = W \oplus \mathbb{R} \cdot X$.

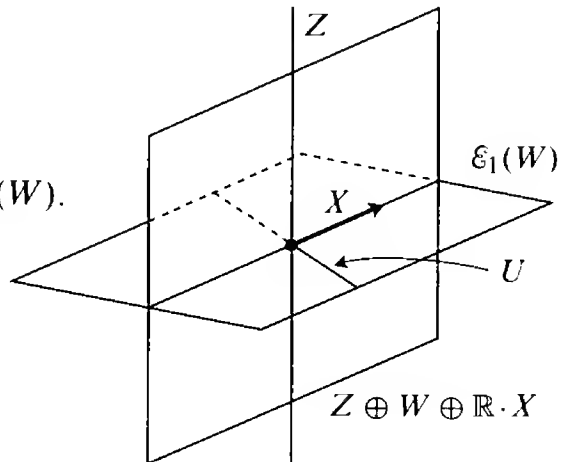


All such submanifolds P can be described locally as follows. Choose $Z \subset M_p$ with $\mathcal{E}_{k+1}(W) \oplus Z = M_p$, so that

$$\dim Z = n - \dim \mathcal{E}_{k+1}(W) = c_{k+1}(W).$$

Then $W \oplus \mathbb{R} \cdot X \oplus Z$ has dimension

$$\dim(W \oplus \mathbb{R} \cdot X \oplus Z) = k + 1 + c_{k+1}(W).$$



Also choose $U \subset M_p$ with $W \oplus \mathbb{R} \cdot X \oplus U = \mathcal{E}_{k+1}(W)$, so that

$$\dim U = \dim \mathcal{E}_{k+1}(W) - k - 1.$$

Then P can be written as the graph of a function from $W \oplus \mathbb{R} \cdot X \oplus Z$ to U . In classical terminology, the submanifolds P , and hence the desired integral manifolds N' , “depend on $\dim \mathcal{E}_{k+1}(W) - k - 1$ arbitrary functions of $k + 1 + c_{k+1}(W)$ variables”.

To prove that N' exists, we can assume without loss of generality that $M = \mathbb{R}^n$, with $p = 0 \in \mathbb{R}^n$, and that

$$\begin{aligned} (e_1)_0, \dots, (e_k)_0 &\text{ is a good basis for } W \\ X &\text{ is } (e_{k+1})_0 \\ \mathcal{E}_{k+1}(W) &\text{ is spanned by } (e_1)_0, \dots, (e_k)_0, (e_{k+1})_0, \dots, (e_l)_0 \\ Z &\text{ is spanned by } (e_{l+1})_0, \dots, (e_n)_0 \\ U &\text{ is spanned by } (e_{k+2})_0, \dots, (e_l)_0. \end{aligned}$$

By Theorem I.2-10(2), we can assume, by composing \mathbb{R}^n with a diffeomorphism, that

$$(1) \quad P = \{(x^1, \dots, x^{k+1}, 0, \dots, 0, x^{l+1}, \dots, x^n)\}.$$

Let N be

$$N = \{(x^1, \dots, x^k, f_{k+1}(x^1, \dots, x^k), \dots, f_n(x^1, \dots, x^k))\},$$

for certain functions f_{k+1}, \dots, f_n with

$$(2) \quad D_i f_t = 0 \quad i = 1, \dots, k; \quad t = k + 1, \dots, n.$$

In order to have $P \supset N$ we must have

$$\begin{aligned} f_{k+1}(x^1, \dots, x^k) &= x^{k+1} \\ f_{k+2}(x^1, \dots, x^k) &= \dots = f_l(x^1, \dots, x^k) = 0 \\ f_v(x^1, \dots, x^k) &= x^v \quad v = l + 1, \dots, n. \end{aligned}$$

Now the map

$$(x^1, \dots, x^n) \mapsto (x^1, \dots, x^l, x^{l+1} - f_{l+1}(x^1, \dots, x^k), \dots, x^n - f_n(x^1, \dots, x^k))$$

has Jacobian matrix equal to the identity at 0, by (2); so by another application of Theorem I.2-10(2) we can assume that

$$(3) \quad N = \{(x^1, \dots, x^k, 0, \dots, 0)\}.$$

The required N' must be of the form

$$(4) \quad N' = \{(x^1, \dots, x^{k+1}, 0, \dots, 0, g_{l+1}(x^1, \dots, x^{k+1}), \dots, g_n(x^1, \dots, x^{k+1}))\},$$

where the functions g_v satisfy

$$(5) \quad g_v(x^1, \dots, x^k, 0) = 0 \quad v = l+1, \dots, n.$$

If

$$\omega = \sum_{i_1 < \dots < i_{k+1}} \omega_{i_1 \dots i_{k+1}} dx^{i_1} \wedge \dots \wedge dx^{i_{k+1}}$$

is any $(k+1)$ -form, then ω restricted to N' is zero if and only if the coefficient of $dx^1 \wedge \dots \wedge dx^{k+1}$ is zero when we replace

$$\begin{aligned} \omega_{i_1 \dots i_{k+1}} & \text{ by} \\ (x^1, \dots, x^{k+1}) & \mapsto \omega_{i_1 \dots i_{k+1}}(x^1, \dots, x^{k+1}, 0, \dots, 0, g_{l+1}(x^1, \dots, x^{k+1}), \\ & \dots, g_n(x^1, \dots, x^{k+1})) \end{aligned}$$

$$dx^j \text{ by } 0 \quad j = k+2, \dots, l$$

$$dx^v \text{ by } \sum_{i=1}^{k+1} \frac{\partial g_v}{\partial x^i} dx^i \quad v = l+1, \dots, n.$$

Thus ω restricted to N' is zero if and only if

$$\begin{aligned} & \sum_{\mu=l+1}^n \omega_{12 \dots p \mu}(x^1, \dots, x^{k+1}, 0, \dots, 0, g_{l+1}(x^1, \dots, x^{k+1}), \\ & \dots, g_n(x^1, \dots, x^{k+1})) \frac{\partial g_\mu}{\partial x^{k+1}} \\ & = \text{certain terms involving the } \partial g_\rho / \partial x^h, \quad h \leq k. \end{aligned}$$

We can write this as

$$\begin{aligned} (6) \quad & \sum_{\mu=l+1}^n C_\mu(x^1, \dots, x^{k+1}, g_{l+1}, \dots, g_n) \cdot \frac{\partial g_\mu}{\partial x^{k+1}} \\ & = D \left(x^1, \dots, x^{k+1}, \dots, g_\rho, \dots, \frac{\partial g_\rho}{\partial x^h}, \dots \right) \\ & \quad [\text{all } g_\rho \text{ and } \partial g_\rho / \partial x^h \text{ evaluated at } (x^1, \dots, x^{k+1})], \end{aligned}$$

where

$$(7) \quad C_\mu(x^1, \dots, x^{k+1}, g_{l+1}, \dots, g_n) \\ = \omega_{12 \dots p\mu}(x^1, \dots, x^{k+1}, 0, \dots, 0, g_{l+1}, \dots, g_n).$$

Choose $n - l$ ($k + 1$)-forms

$$\omega^{l+1}, \dots, \omega^n \in \mathfrak{L}_{k+1}$$

so that, with the conventions of the proof of Lemma 14, the $n - l$ linear functions

$$Y \mapsto \omega^{(v)}(0, e_1, \dots, e_k, Y)$$

are linearly independent. This means that

$$0 \neq \det(\omega^{(v)}(0, e_1, \dots, e_k, e_\mu)) \quad l + 1 \leq \mu, v \leq n.$$

So if we write $\omega^{(v)}$ as

$$\omega^{(v)} = \sum_{i_1 < \dots < i_{k+1}} \omega_{i_1 \dots i_{k+1}}^{(v)} dx^{i_1} \wedge \dots \wedge dx^{i_{k+1}},$$

then

$$(8) \quad 0 \neq \det(\omega_{12 \dots k\mu}^{(v)}(0)) \quad l + 1 \leq \mu, v \leq n.$$

Consider the equations (6) for each $\omega^{(v)}$:

$$(9) \quad \sum_{\mu=l+1}^n C_\mu^{(v)}(x^1, \dots, x^{k+1}, g_{l+1}, \dots, g_n) \frac{\partial g_\mu}{\partial x^{k+1}} \\ = D \left(x^1, \dots, x^{k+1}, \dots, g_\rho, \dots, \frac{\partial g_\rho}{\partial x^h}, \dots \right).$$

Equation (7), together with (8), shows that

$$0 \neq \det(C_\mu^{(v)}(0)) \quad l + 1 \leq \mu, v \leq n.$$

So equations (9) can be written, near 0, as

$$(10) \quad \frac{\partial g_v}{\partial x^{k+1}} = E_v \left(x^1, \dots, x^{k+1}, \dots, g_\rho, \dots, \frac{\partial g_\rho}{\partial x^h}, \dots \right).$$

Now we have arrived at a familiar looking problem.

15. **THEOREM (THE CARTAN-KÄHLER THEOREM).** Let M be an analytic manifold, and let \mathcal{I} be a differential system (of analytic forms) with $\mathcal{I}_0 = 0$. Let $W \subset M_p$ be a regular k -dimensional integral element, and let N be a k -dimensional integral submanifold of \mathcal{I} with $N_p = W$. Let $X \in \mathcal{E}_{k+1}(W)$ be a vector not in W , and let P be an analytic submanifold of M of dimension $k + 1 + c_{k+1}(W)$ such that $P \supset N$ and $P_p \cap \mathcal{E}_{k+1}(W) = W \oplus \mathbb{R} \cdot X$. Then there is a unique analytic $(k + 1)$ -dimensional integral submanifold N' of \mathcal{I} with $N \subset N' \subset P$ and $N'_p = W \oplus \mathbb{R} \cdot X$.

PROOF. The previous considerations show that the existence of N' is equivalent to the existence of functions g_ν satisfying

$$g_\nu(x^1, \dots, x^k, 0) = 0 \quad \nu = l + 1, \dots, n$$

and also equations (6) for all $\omega \in \mathcal{I}_{k+1}$. In particular, the functions g_ν must satisfy (10), with the above initial conditions. The Cauchy-Kowalewski theorem (together with the considerations at the end of section 3) shows that there are unique analytic functions g_ν with this property. This already proves uniqueness, and proves the existence of N' , with inclusion map $i: N' \rightarrow \mathbb{R}^n$, satisfying

$$i^* \omega^{l+1} = \dots = i^* \omega^n = 0.$$

To complete the proof of existence we must show that $i^* \omega = 0$ for all $\omega \in \mathcal{I}_{k+1}$. Here is where the regularity of W is required.

We will continue to use the convention in the proof of Lemma 14. For each $h \leq k$, choose $(h + 1)$ -forms $\omega_{h+1}^{(\alpha)}$ such that the linear functions

$$Y \mapsto \omega_{h+1}^{(\alpha)}(0, e_1, \dots, e_h, Y) \quad \alpha = 1, \dots, c_{h+1} = c_{h+1}(0, e_1, \dots, e_h)$$

are linearly independent. Thus the forms $\omega_{k+1}^{(\alpha)}$ are the forms $\omega^{l+1}, \dots, \omega^n$ introduced previously. Consider the $(k + 1)$ -forms

$$(I) \quad \left\{ \begin{array}{ll} \omega_1^{(\alpha)} \wedge dx^2 \wedge dx^3 \wedge \dots \wedge dx^{k+1} & \alpha = 1, \dots, c_1 \\ \omega_2^{(\alpha)} \wedge dx^3 \wedge \dots \wedge dx^{k+1} & \alpha = 1, \dots, c_2 \\ \vdots & \\ \omega_{k+1}^{(\alpha)} & \alpha = 1, \dots, c_{k+1}. \end{array} \right.$$

We use all of these forms to construct a map $G: \mathbb{R}^{n(k+2)} \rightarrow \mathbb{R}^{c_1 + \dots + c_{k+1}}$, defined by

$$G(q, Y_1, \dots, Y_{k+1}) = \left((\omega_1^{(1)} \wedge dx^2 \wedge \dots \wedge dx^{k+1})(q, Y_1, \dots, Y_{k+1}), \right. \\ \left. \dots, \omega_{k+1}^{(c_{k+1})}(q, Y_1, \dots, Y_{k+1}) \right).$$

Since

$$\omega_i^{(\alpha)} \in \mathcal{I}_i \implies (\omega_i^{(\alpha)} \wedge dx^{i+1} \wedge \dots \wedge dx^{k+1})(0, e_1, \dots, e_i, \dots, Y_h, \dots, e_{k+1}) = 0,$$

the Jacobian matrix of G has the form

$$\begin{array}{c} c_1 \{ \\ c_2 \{ \\ \vdots \\ c_{k+1} \{ \end{array} \begin{pmatrix} q & Y_1 & Y_2 & Y_3 & \dots & Y_{k+1} \\ \boxed{A_1} & 0 & 0 & \dots & 0 \\ & \boxed{A_2} & 0 & \dots & 0 \\ & & \boxed{A_3} & & \vdots \\ & & & \ddots & 0 \\ & & & & \boxed{A_{k+1}} \end{pmatrix} \quad \text{at } (0, e_1, \dots, e_{k+1}).$$

By our choice of the $\omega_{h+1}^{(\alpha)}$, the block A_{h+1} has rank c_{h+1} . So the whole matrix has maximal rank $c_1 + \dots + c_{k+1}$. Thus $G^{-1}(0)$ is an (analytic) submanifold of $\mathbb{R}^{c_1 + \dots + c_{k+1}}$ near $(0, e_1, \dots, e_{k+1})$, of dimension $n(k+2) - c_1 - \dots - c_{k+1}$. But the forms (I) are all in the ideal \mathcal{I} , so $G^{-1}(0)$ contains the manifold $\tilde{\mathcal{M}}_{k+1}$ in the proof of Lemma 14. It also has the same dimension as this manifold, so it *equals* this manifold near $(0, e_1, \dots, e_{k+1})$. We will write the forms in (I) as

$$(II) \quad \begin{cases} \eta^{(\beta)} \wedge dx^{k+1} & \beta = 1, \dots, d = c_1 + \dots + c_k \\ \omega_{k+1}^{(\alpha)} & \alpha = 1, \dots, c_{k+1}, \end{cases}$$

where the forms $\eta^{(\beta)}$ are all in \mathcal{I}_k .

Now consider an arbitrary $(k+1)$ -form $\omega \in \mathcal{I}_{k+1}$. Since W is a regular integral element, we know that we can write

$$\omega(q, Y_1, \dots, Y_k, Y) = \sum_{\alpha=1}^{c_{k+1}} B_{\alpha}(q, Y_1, \dots, Y_k) \cdot \omega_{k+1}^{(\alpha)}(q, Y_1, \dots, Y_k, Y)$$

for all $(q, Y_1, \dots, Y_k) \in \tilde{\mathcal{M}}_k$ close to $(0, e_1, \dots, e_k)$. The functions B_{α} can be solved for explicitly by Cramer's rule, so they are actually analytic functions in a whole neighborhood of $(0, e_1, \dots, e_k)$ in $\mathbb{R}^{n(k+1)}$, even though the equation need hold only for $(q, Y_1, \dots, Y_k) \in \tilde{\mathcal{M}}_k$. We may express this situation as

follows:

$$\left\{ \begin{array}{l} \text{the function} \\ \text{(a)} \quad \omega - \sum_{\alpha=1}^{c_{k+1}} B_{\alpha} \omega^{(\alpha)} \\ \text{on } \mathbb{R}^{n(k+2)} \text{ vanishes on the submanifold } \tilde{\mathcal{M}}_k \times \mathbb{R}^n, \text{ and hence on the} \\ \text{(analytic) submanifold } \tilde{\mathcal{M}}_{k+1}, \text{ defined by the equations} \\ \text{(b)} \quad \eta^{(\beta)} \wedge dx^{k+1} = 0, \quad \omega_{k+1}^{(\alpha)} = 0. \end{array} \right.$$

It follows easily (Problem 1) that locally the function (a) is a sum of analytic functions times the functions in (b). Consequently, we can write

$$\begin{aligned} \omega(q, Y_1, \dots, Y_{k+1}) &= \sum_{\alpha=1}^{c_{k+1}} C_{\alpha}(q, Y_1, \dots, Y_{k+1}) \cdot \omega^{(\alpha)}(q, Y_1, \dots, Y_{k+1}) \\ &\quad + \sum_{\beta=1}^d D_{\beta}(q, Y_1, \dots, Y_{k+1}) \cdot (\eta^{(\beta)} \wedge dx^{k+1})(q, Y_1, \dots, Y_{k+1}), \end{aligned}$$

for analytic C_{α} and D_{β} . This implies that if $q \in N'$, and Y_1, \dots, Y_{k+1} are tangent to N' , then

$$\begin{aligned} (i^* \omega)(q, Y_1, \dots, Y_{k+1}) &= 0 + \sum_{\beta=1}^d D_{\beta}(q, Y_1, \dots, Y_{k+1}) i^*(\eta^{(\beta)} \wedge dx^{k+1})(q, Y_1, \dots, Y_{k+1}). \end{aligned}$$

So it suffices to show that

$$i^*(\eta^{(\beta)} \wedge dx^{k+1}) = 0 \quad \beta = 1, \dots, d.$$

From the form of N' (equation (4) on page 119) it is clear that x^1, \dots, x^{k+1} is a coordinate system on N' . So we write each $i^* \eta^{(\beta)}$ as

$$i^* \eta^{(\beta)} = \sum_{j=1}^{k+1} (-1)^{j+1} h_j^{(\beta)} dx^1 \wedge \dots \wedge \widehat{dx^j} \wedge \dots \wedge dx^{k+1}.$$

Now the above analysis for the $(k+1)$ -form $\omega \in \mathfrak{L}_{k+1}$ can be applied, in particular, for $\omega = \eta^{(\beta)} \wedge dx^j$. Thus each $i^*(\eta^{(\beta)} \wedge dx^j)$ is a linear combination, with analytic coefficients, of the forms $i^*(\eta^{(\beta)} \wedge dx^{k+1})$. Since

$$\begin{aligned} i^*(dx^j \wedge \eta^{(\beta)}) &= dx^j \wedge i^*\eta^{(\beta)} = h_j^{(\beta)} dx^1 \wedge \cdots \wedge dx^{k+1} & j \leq k \\ i^*(dx^{k+1} \wedge \eta^{(\beta)}) &= h_{k+1}^{(\beta)} dx^1 \wedge \cdots \wedge dx^{k+1} \\ &= H^{(\beta)} dx^1 \wedge \cdots \wedge dx^{k+1}, \quad \text{say,} \end{aligned}$$

this shows that we can write $h_j^{(\beta)}$ for $j \leq k$ as an analytic linear combination of the $H^{(\beta)}$,

$$h_j^{(\beta)} = \sum_{\gamma=1}^d E_{j\beta\gamma} H^{(\gamma)}.$$

Since $d\mathfrak{L} \subset \mathfrak{L}$, we can also write each $i^*d\eta^{(\beta)}$ as a linear combination of the $i^*(\eta^{(\beta)} \wedge dx^{k+1})$,

$$i^*d\eta^{(\beta)} = \sum_{\gamma=1}^d F_{\beta\gamma} H^{(\gamma)} dx^1 \wedge \cdots \wedge dx^{k+1}.$$

But

$$\begin{aligned} i^*d\eta^{(\beta)} &= di^*\eta^{(\beta)} \\ &= d\left(\sum_{j=1}^{k+1} (-1)^{j+1} h_j^{(\beta)} dx^1 \wedge \cdots \wedge \widehat{dx^j} \wedge \cdots \wedge dx^{k+1}\right) \\ &= \left[\frac{\partial h_1^{(\beta)}}{\partial x^1} + \cdots + \frac{\partial h_{k+1}^{(\beta)}}{\partial x^{k+1}}\right] dx^1 \wedge \cdots \wedge dx^{k+1} \\ &= \left[\frac{\partial(\sum_{\gamma} E_{1\beta\gamma} H^{(\gamma)})}{\partial x^1} + \cdots + \frac{\partial(\sum_{\gamma} E_{k\beta\gamma} H^{(\gamma)})}{\partial x^k} + \frac{\partial H^{(\beta)}}{\partial x^{k+1}}\right] dx^1 \wedge \cdots \wedge dx^{k+1}. \end{aligned}$$

Comparing with the original expression for $i^*d\eta^{(\beta)}$, we see that we have a system of equations

$$(*) \quad \frac{\partial H^{(\beta)}}{\partial x^{k+1}} = \sum_{\gamma=1}^d F_{\beta\gamma} H^{(\gamma)} + \sum_{j=1}^k G_{j\beta\gamma} \frac{\partial H^{(\gamma)}}{\partial x^j}, \quad \beta = 1, \dots, d,$$

with everything in sight being analytic. Finally, we have to use the fact that the original manifold N (equation (3) on page 119) is an integral submanifold of \mathcal{L} . This implies that all forms $\eta^{(\beta)}$ vanish on N , which means that

$$(*) \quad H^{(\beta)}(x^1, \dots, x^k, 0) = 0, \quad \beta = 1, \dots, d.$$

The uniqueness part of the Cauchy-Kowalewski theorem shows that the only solutions $H^{(\beta)}$ of $(*)$ with the initial conditions $(*_0)$ is $H^{(\beta)} = 0$. Thus all $h_j^{(\beta)} = 0$, so all $i^*\eta^{(\beta)} = 0$. ♦

As an immediate consequence we obtain

16. COROLLARY. Let M be an analytic manifold, and let \mathcal{L} be a differential system (of analytic forms) with $\mathcal{L}_0 = 0$. Let $W \subset M_p$ be a k -dimensional integral element which contains a regular $(k - 1)$ -dimensional integral element. Then there is a k -dimensional analytic integral submanifold N of \mathcal{L} with $N_p = W$.

PROOF. Choose a good basis X_1, \dots, X_k of W , and consider the subspaces $W_1 \subset W_2 \subset \dots \subset W_k$ with W_i the subspace spanned by X_1, \dots, X_i . The desired result then follows by induction from Theorem 15, starting with p as a 0-dimensional integral submanifold. ♦

The reader may easily check that if \mathcal{L} is an ideal generated by linearly independent 1-forms $\omega_1, \dots, \omega_l$, then for every k -dimensional integral element W we have $c_{k+1}(W) = n - l$. Consequently, every integral element is regular. Thus the Frobenius theorem follows, in the analytic case, from the Cartan-Kähler theorem.

As a final remark, we point out that it is not hard to take care of the case $\mathcal{L}_0 \neq 0$. One merely has to assume that $\{q \in M : f(q) = 0 \text{ for all } f \in \mathcal{L}_0\}$ is a submanifold $M' \subset M$ near p , and then apply the previous considerations to $\mathcal{L}|_{M'}$.

ADDENDUM 2

AN ELEMENTARY MAXIMUM PRINCIPLE

It is well-known that if u is harmonic ($\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0$), then u cannot have a relative maximum at an interior point of an open set. A more general principle holds, and its proof, although tricky, is elementary.

On an open set $U \subset \mathbb{R}^n$, consider the second order differential operator L defined by

$$(*) \quad Lu = \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^n b_i \frac{\partial u}{\partial x_i} + cu,$$

for certain functions a_{ij}, b_i, c on U . We assume that $a_{ij} = a_{ji}$, and that the matrix $A = (a_{ij})$ is everywhere definite. [Thus the equation $Lu = 0$ is the most general second order linear elliptic equation.] To be more specific, we will assume that $A = (a_{ij})$ is *positive* definite. Thus $\sum_{i,j} a_{ij} \xi_i \xi_j > 0$ for $0 \neq \xi \in \mathbb{R}^n$; equivalently, if t denotes the transpose, the 1×1 matrix

$$\xi \cdot A \cdot \xi^t > 0 \quad \text{for } 0 \neq \xi \in \mathbb{R}^n.$$

An elementary observation about definite matrices will be needed. Suppose that B is also positive definite, so that

$$\xi \cdot B \cdot \xi^t > 0 \quad \text{for } 0 \neq \xi \in \mathbb{R}^n.$$

For any non-singular matrix P we then have

$$\xi \cdot PBP^t \cdot \xi^t = (\xi P)B(\xi P)^t > 0 \quad \text{for } 0 \neq \xi \in \mathbb{R}^n,$$

so $PBP^t = C = (c_{ij})$ is also positive definite. Now the symmetric matrix A can be diagonalized—there is an orthogonal matrix P such that

$$PAP^{-1} = PAP^t = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad \lambda_i > 0.$$

Then

$$\begin{aligned} \text{trace } AB &= \text{trace } PABP^t = \text{trace}(PAP^t)(PBP^t) \\ &= \text{trace} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} C \\ &= \text{trace}(\lambda_i c_{ij}) \\ &= \sum_i \lambda_i c_{ii} > 0. \end{aligned}$$

Similarly, we have $\text{trace } AB \geq 0$ if B is positive semi-definite, and $\text{trace } AB \leq 0$ if B is negative semi-definite.

Now consider the operator $(*)$, where we assume that

$$(i) \quad c \leq 0 \quad \text{in } U.$$

Suppose that $u: U \rightarrow \mathbb{R}$ is a twice differentiable function with a relative maximum at some point $p \in U$. Assume, moreover, that

$$(ii) \quad u(p) \geq 0.$$

From (i) and (ii) we have

$$(iii) \quad \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}(p) = (Lu)(p) - c(p)u(p) \geq Lu(p).$$

On the other hand, since u has a relative maximum at p , the matrix

$$B = \left(\frac{\partial^2 u}{\partial x_i \partial x_j}(p) \right)$$

is negative semi-definite. Hence we have

$$(iv) \quad 0 \geq \text{trace } A(p) \cdot B = \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}(p).$$

Since (iii) and (iv) imply that $Lu(p) \leq 0$, we find

- (A) If the operator $(*)$ has (a_{ij}) positive definite on U and $c \leq 0$ on U , and the twice differentiable function u satisfies $Lu > 0$ on U , then u cannot have a non-negative relative maximum on U .

The significant fact is that we can replace the condition $Lu > 0$ by $Lu \geq 0$, provided that we consider actual maxima rather than relative maxima.

17. THEOREM (E. HOPF). Consider a second order differential operator

$$Lu = \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^n b_i \frac{\partial u}{\partial x_i} + cu \quad c \leq 0$$

on a connected open set $U \subset \mathbb{R}^n$. Assume that the functions b_i and c are locally bounded, and that in a neighborhood of any point of U there are constants $\varepsilon, M > 0$ such that the matrix (a_{ij}) satisfies

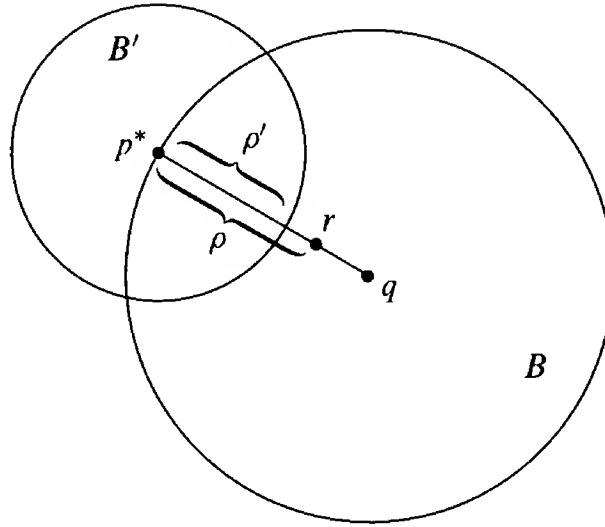
$$\varepsilon \cdot \sum_{i=1}^n \xi_i^2 \leq \sum_{i,j=1}^n a_{ij} \xi_i \xi_j \leq M \cdot \sum_{i=1}^n \xi_i^2 \quad \xi \in \mathbb{R}^n.$$

Suppose that u is a twice differentiable function on U satisfying

$$Lu \geq 0.$$

Then u cannot have a non-negative maximum on U , unless u is a constant.

PROOF. Suppose u has a maximum at $p \in U$, with $u(p) \geq 0$. If u is not constant, then there is clearly a point $q \in U$ and an open ball B centered at q with $\bar{B} \subset U$ such that $u(q) < u(p)$, but $u(p^*) = u(p)$ for some $p^* \in$ boundary B . Moreover, by choosing the smallest ball B with this property, we



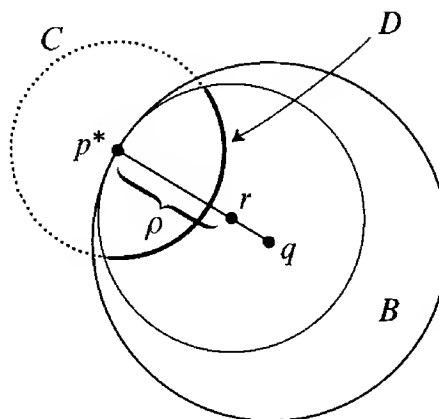
can assume that $u < u(p)$ in B . Let r be a point on the open segment $\overline{qp^*}$, set

$$\rho = d(r, p^*) \quad \text{and choose} \quad 0 < \rho' < \rho.$$

Let B' be the open ball of radius ρ' around p^* ; assume ρ' chosen sufficiently small so that $\bar{B'} \subset U$.

Now consider the function

$$v(x) = e^{-kd(r,x)^2} - e^{-k\rho^2}$$

$$Lv(x) = e^{-kd(r,x)^2} \left[4k^2 \sum_{i,j} a_{ij} (x_i - r_i)(x_j - r_j) - 2k \sum_i b_i (x_i - r_i) \right] + c \left[e^{-k\rho^2} - e^{-kd(r,x)^2} \right].$$
$$(1) \quad L(u + \lambda v) = Lu + \lambda Lv > 0 \quad \text{in } \overline{B'}.$$
$$(2) \quad v(p^*) = 0 \implies (u + \lambda v)(p^*) = u(p^*) = u(p).$$
$$\begin{aligned} x \in C &\implies d(x, r) > \rho \\ &\implies v(x) < 0 \\ x \in D &\implies x \in B \implies u(x) < u(p). \end{aligned}$$

$$(3) \quad u + \lambda v < u(p) \quad \text{on } S'.$$

The example

$$u = -(x^2 + y^2) - 4$$
$$Lu = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - u = x^2 + y^2 \geq 0$$

shows that the function u in Theorem 17 may well have a negative maximum on U . The example

$$u = -(x^2 + y^2) + 5$$

$$Lu = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + u = 1 - (x^2 + y^2) \geq 0 \quad \text{for } x^2 + y^2 \leq 1$$

shows that the hypothesis $c \leq 0$ is essential. If we assume $c = 0$, then we get a stronger conclusion:

18. COROLLARY. Consider the operator L of Theorem 17, with $c = 0$. If u is a twice differentiable function on U with $Lu \geq 0$, then u cannot have a maximum on U unless u is a constant.

PROOF. Suppose u has a maximum at p . Let $v = u - u(p)$. Then v has a maximum of 0 at p . Moreover,

$$Lv = Lu \geq 0.$$

So Theorem 17 implies that v is a constant. ♦

As an application, consider a function $f: M \rightarrow \mathbb{R}$ on a Riemannian manifold M . Then we have the Laplacian Δf , defined in Addendum 1 to Chapter 7. In a coordinate system (x^1, \dots, x^n) on M , the formula for Δf (pg. IV.133) is precisely of the form considered in Corollary 18. So if $\Delta f \geq 0$, then Δf cannot have a maximum on M , unless f is a constant function. This gives another proof of Bochner's Lemma (Lemma 7-60).

In contrast to Corollary 18, where we assume $c = 0$, there is another result where c is arbitrary.

19. COROLLARY. Consider the operator L of Theorem 17, with arbitrary c . If u is a twice differentiable function on U with $Lu \geq 0$ and $u \leq 0$, then u cannot have the value 0 anywhere on U unless u is identically 0.

PROOF. Let

$$Pu = \sum_{i,j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^n b_i \frac{\partial u}{\partial x_i}.$$

Then we have

$$Pu + cu = Lu \geq 0.$$

Hence

$$Pu + \min(c, 0)u = [\min(c, 0) - c]u + Lu.$$

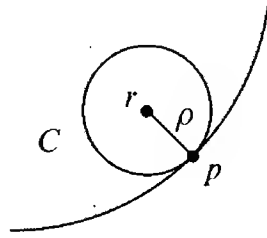
Now

$$\begin{aligned} \min(c, 0) &\leq 0 \\ \min(c, 0) - c &\leq 0 \implies [\min(c, 0) - c]u \geq 0 \quad \text{since } u \leq 0 \\ &\implies [\min(c, 0) - c]u + Lu \geq 0. \end{aligned}$$

Applying Theorem 17 to the operator $Pu + \min(c, 0)u$, we conclude that u cannot have a non-negative maximum unless it is a constant. ♦

There is also a version of Theorem 17 when u has its maximum at a boundary point of U .

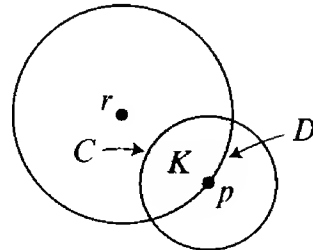
20. THEOREM. Consider a second order differential operator L as in Theorem 17. Let u be a twice differentiable function on U satisfying $Lu \geq 0$ and such that at some point $p \in \text{boundary } U$ the function u has a maximum $u(p) \geq 0$ on $U \cup \{p\}$. Suppose moreover that there is some closed ball $\overline{B_\rho(r)} \subset U \cup \{p\}$ containing p on its boundary, and that the directional derivative of u at p in the



direction from p to r is ≥ 0 (all directional derivatives in directions tangent to the boundary of $B_\rho(r)$ are clearly equal to 0). Assume, finally, that the functions a_{ij}, b_i, c have the same properties as in Theorem 17, but in $U \cup \{p\}$. Then u is a constant function.

PROOF. Suppose u is not a constant function. Choose $0 < \rho_1 < \rho = d(p, r)$, and let

$$K = \{x : d(x, r) \leq \rho \text{ and } d(x, p) \leq \rho_1\}.$$



Again define

$$v(x) = e^{-kd(r,x)^2} - e^{-k\rho^2}.$$

The boundary of K is the union of a closed set $C \subset U$ and a closed set D on which $v = 0$. Theorem 17 implies (choosing ρ smaller if necessary) that $u(x) < u(p)$ for $x \in C$. So for sufficiently small $\lambda > 0$ we have

$$-\lambda v(x) \geq u(x) - u(p) \quad \text{for } x \in \text{boundary } K.$$

But for sufficiently large k , we have $Lv > 0$ in K . So

$$L(u - u(p) + \lambda v) = Lu - cu(p) + \lambda Lv > 0.$$

It follows from (A) that

$$-\lambda v(x) \geq u(x) - u(p) \quad \text{for all } x \in K.$$

So the directional derivative of u at p , in the direction from p to r , is less than or equal to this directional derivative at p of $-\lambda v$. But we easily compute that the latter directional derivative is

$$-2\lambda k \rho e^{-k\rho^2} < 0,$$

contradicting the hypotheses. ♦

Naturally, there are analogous versions of Corollaries 18 and 19.

PROBLEM

1. (a) Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a C^∞ [respectively, C^ω] function such that $f = 0$ on the points $(0, \dots, 0, x^{n-k+1}, \dots, x^n)$ near 0. Show that there are C^∞ [C^ω] functions h_i near 0 such that

$$f = \sum_{i=1}^k h_i x^i.$$

[The C^ω case is actually trivial; the C^∞ case can be proved by generalizing the argument in the proof of Lemma I.3-2.]

- (b) Let $g: \mathbb{R}^n \rightarrow \mathbb{R}^k$ be a C^∞ [C^ω] function whose Jacobian has rank k on $g^{-1}(0)$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a C^∞ [C^ω] function which vanishes on $g^{-1}(0)$. Then near any point of $g^{-1}(0)$ there are C^∞ [C^ω] functions h_i such that

$$f = \sum_{i=1}^k h_i g^i.$$

CHAPTER 11

EXISTENCE AND NON-EXISTENCE OF ISOMETRIC IMBEDDINGS

In the past we have had some very special results about the non-existence of isometric imbeddings of certain Riemannian manifolds in other Riemannian manifolds. For example, a compact surface of everywhere negative curvature cannot be isometrically imbedded, or even immersed, in \mathbb{R}^3 , nor can a complete surface of constant negative curvature be isometrically immersed in \mathbb{R}^3 . Ideally, differential geometry should be replete with such results, so that we could have a reasonable chance of finding the smallest dimensional Euclidean space into which a given Riemannian manifold can be isometrically imbedded. But at present only quite isolated facts are known, and a general theory can hardly be said to exist.

There are, first of all, purely topological, or at any rate differential-topological, questions which have to be considered in any imbedding problem—for there is no point trying to isometrically immerse or imbed a Riemannian manifold in \mathbb{R}^m unless its underlying differentiable manifold has some differentiable immersion or imbedding in \mathbb{R}^m . Generally speaking, the methods used to settle such questions are of little interest to differential geometry *per se*. We note, however, that one special result of this sort has already been proved in Volume I: A compact hypersurface imbedded in \mathbb{R}^m is always orientable (Theorem I.11-14). Thus, for example, there is no imbedding of the projective plane \mathbb{P}^2 in \mathbb{R}^3 . We can supplement this result with a simple differential geometric one: If $\langle \ , \ \rangle$ is a metric on \mathbb{P}^2 with $K > 0$, then $(\mathbb{P}^2, \langle \ , \ \rangle)$ cannot even be isometrically immersed in \mathbb{R}^3 ; this follows directly from Hadamard's Theorem (Theorem 2-11).

At the other extreme from these global topological restrictions, there are certain purely local results. For example, if $n \geq 3$, and M^n has all sectional curvatures < 0 , then M^n cannot be locally isometrically imbedded in \mathbb{R}^{n+1} ; for the principal curvatures k_1, \dots, k_n would have to satisfy $k_i k_j < 0$ for all i, j , while some pair k_i, k_j must have the same sign. Similarly, Theorem 7-50 shows that if $n \geq 3$, and $M^n \subset \mathbb{R}^{n+1}$ has Ricci tensor $\text{Ric} = 0$, then M is flat; so for $n \geq 3$, a non-flat M^n with $\text{Ric} = 0$ is not isometrically imbeddable in \mathbb{R}^{n+1} . Historically, this was first used to show that the 4-dimensional Schwartzschild metric of general relativity is not imbeddable in \mathbb{R}^5 .

To obtain purely local results for higher codimension, we need to replace the trivial algebraic considerations used previously by something more substantial.

1. LEMMA. Let $s: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^k$ be a symmetric bilinear map, and let $\langle \cdot, \cdot \rangle$ be a positive definite inner product on \mathbb{R}^k . Let $S^{n-1} \subset \mathbb{R}^n$ be the unit sphere (with respect to the usual inner product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^n), and consider the function $f(x) = |s(x, x)|^2$ for $x \in S^{n-1}$.

(1) If $x \in S^{n-1}$ is a critical point of f , then

$$\langle s(x, x), s(x, y) \rangle = 0 \quad \text{for all } y \text{ with } \langle x, y \rangle = 0.$$

Consequently, if $f(x) \neq 0$, then

$$s(x, y) = 0 \implies \langle x, y \rangle = 0.$$

(2) If $x \in S^{n-1}$ is a minimum point of f , then for all $y \in S^{n-1}$ with $\langle x, y \rangle = 0$ we have

$$\langle s(x, x), s(y, y) \rangle + 2\langle s(x, y), s(x, y) \rangle \geq \langle s(x, x), s(x, x) \rangle.$$

PROOF. (1) Using the fact that the derivative $DA(x)$ of a linear transformation A is always A itself, we easily see that the map $S: \mathbb{R}^n \rightarrow \mathbb{R}^k$ defined by $S(x) = s(x, x)$ has derivative $DS(x): \mathbb{R}^n \rightarrow \mathbb{R}^k$ given by

$$(DS)(x)(y) = 2s(x, y).$$

It follows that

$$(Df)(x)(y) = 2\langle S(x)(y), S(x) \rangle = 4\langle s(x, y), s(x, x) \rangle.$$

Since $x \in S^{n-1}$ is a critical point for f , we must have $(Df)(x)(y) = 0$ for all $y \in S^{n-1}_x$, i.e., for $\langle x, y \rangle = 0$.

Now suppose $s(x, y) = 0$ and $s(x, x) \neq 0$. Writing $y = \lambda x + y'$, with

$$\langle x, y' \rangle = 0 \implies \langle s(x, x), s(x, y') \rangle = 0 \quad \text{by the above paragraph,}$$

we have

$$\begin{aligned} 0 &= \langle s(x, x), s(x, y) \rangle \\ &= \langle s(x, x), s(x, \lambda x) \rangle + \langle s(x, x), s(x, y') \rangle \\ &= \langle s(x, x), s(x, \lambda x) \rangle. \end{aligned}$$

Since $s(x, x) \neq 0$, this implies that $\lambda = 0$.

(2) Let c be the curve in S^{n-1} defined by

$$c(t) = (\cos t)x + (\sin t)y.$$

Since x is a minimum of f we have

$$0 \leq \left. \frac{d^2}{dt^2} \right|_{t=0} f(c(t)).$$

A short computation shows that the right side is

$$4[-\langle s(x, x), s(x, x) \rangle + \langle s(x, x), s(y, y) \rangle + 2\langle s(x, y), s(x, y) \rangle]. \diamond$$

From this we derive, first of all, a purely local result.

2. PROPOSITION. Let N be a manifold of dimension $2n-2$ with all sectional curvatures $\geq K_0$, and let M be a manifold of dimension n with all sectional curvatures $< K_0$. Then M cannot be isometrically immersed in N . The result also holds if all sectional curvatures of N are $> K_0$ and all sectional curvatures of M are $\leq K_0$.

PROOF. Suppose we could isometrically immerse M in N , and let s be the second fundamental form. For any $p \in M$, and orthonormal $X, Y \in M_p$, Gauss' equation gives, under the first hypothesis,

$$\begin{aligned} K_0 &\leq \langle R'(X, Y)Y, X \rangle \\ &= \langle R(X, Y)Y, X \rangle + \langle s(X, Y), s(X, Y) \rangle - \langle s(X, X), s(Y, Y) \rangle \\ &< K_0 + \langle s(X, Y), s(X, Y) \rangle - \langle s(X, X), s(Y, Y) \rangle. \end{aligned}$$

Under the second hypothesis the \leq and $<$ are interchanged. In either case, we obtain

$$(1) \quad \langle s(X, X), s(Y, Y) \rangle - \langle s(X, Y), s(X, Y) \rangle < 0, \quad \begin{array}{l} X, Y \in M_p \\ \text{linearly independent} \end{array}$$

(for this final inequality we do not need X, Y to be orthonormal). Choose $X \in M_p$ to be a minimum point of $X \mapsto |s(X, X)|$ on the unit sphere of M_p . Since $\{Y : s(X, Y) = 0\}$ has dimension ≥ 2 , there is a unit vector Y linearly independent of X with

$$(2) \quad s(X, Y) = 0.$$

From (1) and (2) we see that we must have $s(X, X) \neq 0$. But then Lemma 1(1) implies that $\langle X, Y \rangle = 0$, so Lemma 1(2) gives

$$\langle s(X, X), s(Y, Y) \rangle + 0 \geq \langle s(X, X), s(X, X) \rangle \geq 0,$$

contradicting (1) and (2). \diamond

In particular, an n -manifold M of constant curvature $K < K_0$ cannot be locally isometrically immersed in a $(2n - 2)$ -manifold N of constant curvature K_0 . An example of a (non-complete) n -manifold of constant negative curvature in \mathbb{R}^{2n-1} is given in Problem 1. It seems reasonable to conjecture that there is no immersion of a complete n -manifold of constant negative curvature in \mathbb{R}^{2n-1} , but this has not been proved (whether one can be found in \mathbb{R}^{2n} is any body's guess). It is known, however, that no such immersion exists if M is compact. This follows from the next theorem, whose proof combines the local information from Lemma 1 with just a smidgen of globalness.

3. PROPOSITION. Let N be the complete simply-connected $(2n - 1)$ -dimensional manifold of constant curvature $K_0 \leq 0$, and let M be a compact n -manifold with all sectional curvatures ≤ 0 . Then M cannot be isometrically immersed in N .

PROOF. Suppose there were an isometric immersion $f: M \rightarrow N$. Let $q_0 \in N$ be a fixed point, and choose $p \in M$ so that $f(p)$ is furthest from q_0 . Then (pg. IV.118) there is $\xi \in M_p^\perp$ with

$$(1) \quad \langle s(X, X), \xi \rangle > \sqrt{-K_0} \implies \langle s(X, X), s(X, X) \rangle > -K_0 \quad \text{for } X \in M_p.$$

Choose $X \in M_p$ to be the minimum point of $X \mapsto |s(X, X)|^2$ on the unit sphere in M_p . Since $\{Y : s(X, Y) = 0\}$ has dimension ≥ 1 , there is a unit vector $Y \in M_p$ with

$$(2) \quad s(X, Y) = 0 \implies \langle X, Y \rangle = 0 \quad \text{by Lemma 1(1).}$$

Then Lemma 1(2) gives

$$(3) \quad \langle s(X, X), s(Y, Y) \rangle \geq \langle s(X, X), s(X, X) \rangle > -K_0, \quad \text{by (1).}$$

Moreover, applying Gauss' equation to the plane $P \subset M_p$ spanned by the orthonormal vectors X, Y , we have

$$\begin{aligned} K(P) &= \langle s(X, X), s(Y, Y) \rangle - \langle s(X, Y), s(X, Y) \rangle + K_0 \\ &> -K_0 + 0 + K_0 \quad \text{by (2) and (3).} \end{aligned}$$

This contradicts the assumption that $K(P) \leq 0$. ♦

Remark: More generally, if each tangent space M_p contains an l -dimensional subspace on which all sectional curvatures are ≤ 0 , and the compact manifold M can be isometrically immersed in the complete simply-connected $(n+k)$ -dimensional manifold of sectional curvature $K_0 \leq 0$, then we must have $k \geq l$. Proposition 2 can be generalized similarly.

4. COROLLARY (TOMPKINS). An n -dimensional compact flat Riemannian manifold cannot be isometrically immersed in \mathbb{R}^{2n-1} .

Obviously $2n - 1$ is the best possible dimension here, since the flat n -torus $S^1 \times \cdots \times S^1$ is isometrically imbedded in \mathbb{R}^{2n} . It is also isometrically imbedded in $H^{2n}(K_0)$ for any $K_0 < 0$, since it is isometrically imbedded in S^{2n-1} , and there are spheres of all curvatures in $H^{2n}(K_0)$. I do not know if there is a non-flat compact n -manifold in \mathbb{R}^{2n} or $H^{2n}(K_0)$ with all sectional curvatures ≤ 0 .

The proof of Proposition 3 breaks down if N is a sphere, with constant curvature $K_0 > 0$, since we cannot guarantee the existence of the requisite point $p \in M$ unless we know that M lies in a hemisphere. Indeed, the n -dimensional flat torus $S^1 \times \cdots \times S^1$ can be isometrically imbedded in S^{2n-1} . It seems reasonable to assume that M^n cannot be isometrically immersed in S^{2n-1} if all sectional curvatures of M are > 0 and < 1 . For the special case where M has constant curvature, see Problem 2.

Now we are going to consider some more elaborate algebraic results. Let V be a real vector space, and let $\Phi: V \times V \rightarrow \mathbb{R}$ be bilinear. With Φ we can associate a linear transformation $\tilde{\Phi}: V \rightarrow V^*$ by

$$\tilde{\Phi}(v)(w) = \Phi(v, w).$$

A collection Φ^1, \dots, Φ^n of bilinear forms on V is called **exteriorly orthogonal** if for all $v_1, v_2 \in V$ we have

$$\sum_{i=1}^n \tilde{\Phi}^i(v_1) \wedge \tilde{\Phi}^i(v_2) = 0 \in \Omega^2(V);$$

equivalently,

$$\sum_{i=1}^n [\Phi^i(v_1, w_1) \Phi^i(v_2, w_2) - \Phi^i(v_1, w_2) \Phi^i(v_2, w_1)] = 0$$

for all $v_1, v_2, w_1, w_2 \in V$. Before stating the main result about exteriorly orthogonal bilinear forms, we make a simple observation. If $\Phi \neq 0$ is $\Phi = \phi \otimes \phi$ for some $\phi \in V^*$, so that $\Phi(v, w) = \phi(v) \cdot \phi(w)$, then $\tilde{\Phi}(v) = \phi(v) \cdot \phi$, and consequently $\text{range } \tilde{\Phi} \subset V^*$ is 1-dimensional. Conversely, if $\text{range } \tilde{\Phi} \subset V^*$ is 1-dimensional, then $\tilde{\Phi}$ must be of the form

$$\tilde{\Phi}(v)(w) = \psi(v) \cdot \phi(w)$$

for $\phi, \psi \in V^*$. If, moreover, $\tilde{\Phi}$ is symmetric, so that $\psi(v) \cdot \phi(w) = \psi(w) \cdot \phi(v)$ for all v, w , then we must have $\phi = \psi$ [since $(\phi - \psi)(w) = 0$ for some $w \neq 0$], so Φ is of the form $\Phi = \phi \otimes \phi$.

5. THEOREM (É. CARTAN). Let V be a real vector space of dimension n , and let Φ^1, \dots, Φ^n be n exteriorly orthogonal symmetric bilinear forms on V . Suppose that

$$(*) \quad 0 = \bigcap_{i=1}^n \ker \tilde{\Phi}^i = \{v \in V : \Phi^i(v, w) = 0 \text{ for all } v \in V \text{ and all } i\}.$$

Then there is an orthogonal $n \times n$ matrix A and linearly independent $\phi^1, \dots, \phi^n \in V^*$ such that

$$\Phi^i = \sum_{j=1}^n A_j^i \phi^j \otimes \phi^j.$$

PROOF. We claim that there is a vector $v \in V$ such that the $\tilde{\Phi}^i(v) \in V^*$ are linearly independent. To prove this, let $v_0 \in V$ be a vector such that the subspace

$$[\tilde{\Phi}^1(v_0), \dots, \tilde{\Phi}^n(v_0)] \subset V^*$$

spanned by the $\tilde{\Phi}^i(v_0)$ has maximal dimension $d \leq n$, and suppose that $d < n$. Replacing the $\{\tilde{\Phi}^i\}$ by an orthogonal linear combination of them changes neither the hypotheses nor the conclusion of the theorem, so without loss of generality we can assume that

$$\begin{aligned} \tilde{\Phi}^1(v_0), \dots, \tilde{\Phi}^d(v_0) &\text{ are linearly independent,} \\ \tilde{\Phi}^{d+1}(v_0) &= \dots = \tilde{\Phi}^n(v_0) = 0. \end{aligned}$$

Then for any vector $v \in V$ we have

$$\sum_{i=1}^d \tilde{\Phi}^i(v_0) \wedge \tilde{\Phi}^i(v) = 0.$$

Cartan's Lemma thus implies that for $i = 1, \dots, d$, the $\tilde{\Phi}^i(v)$ are a linear combination of the $\tilde{\Phi}^i(v_0)$, $i = 1, \dots, d$. Consequently

$$\mathcal{V} = \{\tilde{\Phi}^i(v) : v \in V, 1 \leq i \leq d\} \subset V^*$$

also has dimension exactly d . Since $d < n$, there is a vector $0 \neq w \in V$ such that $\phi(w) = 0$ for all $\phi \in \mathcal{V}$. But by $(*)$ there is some i and some $v \in V$ such that $\Phi^i(v, w) \neq 0$; clearly $i > d$. Now consider the vector $v_0 + \varepsilon v$. If $\varepsilon > 0$ is sufficiently small, then

$$\dim[\tilde{\Phi}^1(v_0 + \varepsilon v), \dots, \tilde{\Phi}^d(v_0 + \varepsilon v)] = \dim[\tilde{\Phi}^1(v_0), \dots, \tilde{\Phi}^d(v_0)] = d$$

$$\Downarrow$$

$$[\tilde{\Phi}^1(v_0 + \varepsilon v), \dots, \tilde{\Phi}^d(v_0 + \varepsilon v)] = \mathcal{V}, \quad \text{since } \dim \mathcal{V} = d.$$

But $\tilde{\Phi}^i(v_0 + \varepsilon v) \notin \mathcal{V}$, by the choice of v and i . So

$$\dim[\tilde{\Phi}^1(v_0 + \varepsilon v), \dots, \tilde{\Phi}^n(v_0 + \varepsilon v)] > d,$$

contradicting the definition of d . This establishes the claim.

Now choose a basis v_1, \dots, v_n of V such that $\tilde{\Phi}^1(v_1), \dots, \tilde{\Phi}^n(v_1)$ are linearly independent. Since

$$\sum_{i=1}^n \tilde{\Phi}^i(v_1) \wedge \tilde{\Phi}^i(v_j) = 0,$$

Cartan's Lemma implies that there is a symmetric matrix $C(j)$, with $C(1) =$ identity, such that

$$\tilde{\Phi}^i(v_j) = \sum_{h=1}^n C(j)_h^i \tilde{\Phi}^h(v_1).$$

The equation

$$\sum_h \tilde{\Phi}^h(v_j) \wedge \tilde{\Phi}^h(v_k) = 0$$

implies that $C(j)$ and $C(k)$ commute. Then a well-known theorem of linear algebra (Problem 3) states that there is an orthogonal matrix B such that the matrices $B \cdot C(i) \cdot B^t$ are diagonal for all i , where t denotes the transpose. If we set

$$\Psi^i = \sum_h B_h^i \Phi^h,$$

then $\tilde{\Psi}^i(v_j)$ is a constant times $\tilde{\Psi}^i(v_1)$. Thus range $\tilde{\Psi}^i$ is 1-dimensional, so $\Psi^i = \phi^i \otimes \phi^i$ for some $\phi^i \in V^*$. We choose $A = B^{-1}$. The ϕ^i must be linearly independent, for otherwise there is $0 \neq v \in V$ such that $\phi^i(v) = 0$ for all i , contradicting (*). ♦

The hypothesis (*) in Theorem 5 may be interpreted as saying that the set $\{\Phi^i\}$ "depends on n variables"—we cannot find $\phi^1, \dots, \phi^{n-1} \in V^*$ such that each Φ^i is a linear combination of the $\phi^j \otimes \phi^k$, $1 \leq j, k \leq n-1$. More generally, if Φ^1, \dots, Φ^k are bilinear forms on V , then the set $\{\Phi^i\}$ **depends on q variables** if the subspace $\bigcap_{i=1}^k \ker \tilde{\Phi}^i$ has codimension q .

6. COROLLARY. Let V be a real vector space of dimension n , and let Φ^1, \dots, Φ^k be k exteriorly orthogonal symmetric bilinear forms on V which depend on

$l \geq k$ variables (so necessarily $k \leq n$). Then $l = k$ and there is an orthogonal $k \times k$ matrix A and linearly independent $\phi^1, \dots, \phi^k \in V^*$ such that

$$\Phi^i = \sum_{j=1}^k A_j^i \phi^j \otimes \phi^j.$$

In particular, k exteriorly orthogonal symmetric bilinear forms always depend on $\leq k$ variables.

PROOF. Without loss of generality, we can assume that $l = n$ [by applying the result to a subspace of V complementary to $\bigcap_{i=1}^k \ker \tilde{\Phi}^i$]. If $k < l = n$, we set

$$\Phi^{k+1} = \dots = \Phi^n = 0.$$

The n bilinear forms Φ^1, \dots, Φ^n are then exteriorly orthogonal and $\bigcap_{i=1}^n \ker \tilde{\Phi}^i = 0$. By the Theorem, there is an orthogonal $n \times n$ matrix A , and linearly independent $\phi^1, \dots, \phi^n \in V^*$ with

$$\Phi^i = \sum_{j=1}^n A_j^i \phi^j \otimes \phi^j.$$

So we cannot have $\Phi^i = 0$ for any i , which shows that actually $k = n = l$. ♦

These algebraic results were used by Cartan for a systematic local study of n -dimensional manifolds M of constant curvature K isometrically imbedded in an $(n + k)$ -dimensional manifold N of constant curvature $K_0 > K$. For an adapted orthonormal moving frame X_1, \dots, X_m on $M \subset N$ we have, as in Chapter 1,

$$\psi_j^r = \sum_i s_{ij}^r \theta^i, \quad s_{ij}^r = s_{ji}^r;$$

the second fundamental forms Π^r are given by

$$\Pi^r = \sum_i \psi_i^r \otimes \theta^i.$$

We also have

$$\begin{aligned} K_0 \theta^i \wedge \theta^j &= \Omega_j^i - \sum_r \psi_i^r \wedge \psi_j^r \\ &= K \theta^i \wedge \theta^j - \sum_r \psi_i^r \wedge \psi_j^r, \end{aligned}$$

or equivalently

$$(1) \quad \sum_r (s_{ij}^r s_{kl}^r - s_{il}^r s_{kj}^r) = (K - K_0)(\delta_{ij}\delta_{kl} - \delta_{il}\delta_{kj}).$$

If we define

$$\Psi = \sqrt{K_0 - K} \left(\sum_i \theta^i \otimes \theta^i \right),$$

then equation (1) says that the $k + 1$ bilinear forms $\{\Pi^r, \Psi\}$ are exteriorly orthogonal. The collection $\{\Pi^r, \Psi\}$ certainly depends on all n variables, since Ψ alone does. So Corollary 6 implies that $n \leq k + 1$, showing once again that M^n cannot be isometrically imbedded in N^{2n-2} . Cartan showed, using his theory of exterior differential systems (Chapter 10, Addendum 1) that the analytic local imbeddings of M^n in N^{2n-1} depend upon $n(n - 1)$ functions of one variable.

Another consequence of Corollary 6 depends on two definitions, one intrinsic and one extrinsic. For a point p of a Riemannian manifold M^n , we define the **index of nullity** at p to be

$$\mu(p) = \dim\{X \in M_p : R(X, Y) = 0 \text{ for all } Y \in M_p\}.$$

Equivalently, $n - \mu(p)$ is the minimum number of 1-forms in terms of which we can express the collection of 2-forms $\{\Omega_j^i(p)\}$. For $M^n \subset \mathbb{R}^{n+k}$, with second fundamental form s , we define the **index of relative nullity** at p to be

$$\begin{aligned} \nu(p) &= \dim\{X \in M_p : A_\xi(X) = 0 \text{ for all } \xi \in M_p^\perp\} \\ &= \dim\{X \in M_p : s(X, Y) = 0 \text{ for all } Y \in M_p\}. \end{aligned}$$

Equivalently, $n - \nu(p)$ is the minimum number of 1-forms in terms of which we can express the collection of forms $\{\Pi^r(p)\}$, for an orthonormal set $\nu^{n+1}, \dots, \nu^{n+k} \in M_p^\perp$.

7. PROPOSITION. For $M^n \subset \mathbb{R}^{n+k}$ we have

$$\nu(p) \leq \mu(p) \leq \nu(p) + \text{rank } s(p) \leq \nu(p) + k.$$

PROOF. The first inequality follows from Gauss' equation, which shows that

$$\begin{aligned} \{X \in M_p : s(X, Y) = 0 \text{ for all } Y \in M_p\} \\ \subset \{X \in M_p : R(X, Y) = 0 \text{ for all } Y \in M_p\}. \end{aligned}$$

For the second we can assume, by choosing the v^r appropriately, that $\Pi^{n+1}(p), \dots, \Pi^{n+d}(p)$ are linearly independent, for $d = \text{rank } s(p)$, while the other $\Pi^r(p)$ are all 0. Then Gauss' equation shows that $\Pi^{n+1}(p), \dots, \Pi^{n+d}(p)$ are exteriorly orthogonal when restricted to $\{X \in M_p : R(X, Y) = 0 \text{ for all } Y \in M_p\}$. Let W be a subspace such that

$$\begin{aligned} \{X \in M_p : R(X, Y) = 0 \text{ for all } Y \in M_p\} \\ = W \oplus \{X \in M_p : s(X, Y) = 0 \text{ for all } Y \in M_p\}. \end{aligned}$$

Then $\Pi^{n+1}(p), \dots, \Pi^{n+d}(p)$ are exteriorly orthogonal on W , and depend on all variables of W . Corollary 6 implies that $d \geq \dim W = v(p) - \mu(p)$. ♦

Remark: We have a similar result for $M^n \subset N^{n+k}$, where N^{n+k} has constant curvature K_0 , provided we redefine

$$\mu(p) = \dim\{X \in M_p : R(X, Y)Z = K_0[\langle Y, Z \rangle X - \langle X, Z \rangle Y]\}.$$

8. COROLLARY. If M^n is a compact manifold immersed in \mathbb{R}^{n+k} , then

$$k \geq \min_{p \in M} \mu(p).$$

PROOF. Proposition 7-30 shows that for some $p \in M$ we have $v(p) = 0$. ♦

Note that Corollary 4 is a special case (admittedly the only reasonably general consequence we can give). Recently Corollary 6 has been used to prove results of quite another sort, which we will mention in the next chapter.

This ends our treatment of non-imbeddability theorems, and pretty much exhausts the subject in its present state (a few other special results are mentioned in the Bibliography). Now we will take a more positive approach to life and try to prove that under certain circumstances isometric imbeddings do exist. We first consider the purely local problem of isometrically imbedding a surface in \mathbb{R}^3 . So we assume that we are given functions g_{ij} ($= E, F, G$) on a neighborhood of $0 \in \mathbb{R}^2$, with $\det(g_{ij}) > 0$, and we want to find a function $f : U \rightarrow \mathbb{R}^3$, on some smaller neighborhood U of $0 \in \mathbb{R}^2$, such that $I_f = f^*(\langle \cdot, \cdot \rangle)$ has components g_{ij} . This means that the component functions f^α of f must satisfy

$$g_{ij} = \sum_{\alpha=1}^3 \langle f^\alpha_i, f^\alpha_j \rangle,$$

so that we have three (non-linear) partial differential equations in three unknowns. We also know that f can be found once we have functions l_{ij} satisfying Gauss' equation and the Codazzi-Mainardi equation, which again gives us three equations in three unknowns. There is also a way of introducing a single second order equation, which was used classically. Suppose that the required f exists; let N be its normal map, and let l_{ij} be the components of Π_f . Then for each component function f^α of f we have the Gauss formulas

$$(1) \quad \begin{aligned} f^\alpha_{11} - \Gamma_{11}^1 f^\alpha_1 - \Gamma_{11}^2 f^\alpha_2 &= l_{11} N^\alpha \\ f^\alpha_{12} - \Gamma_{12}^1 f^\alpha_1 - \Gamma_{12}^2 f^\alpha_2 &= l_{12} N^\alpha \\ f^\alpha_{22} - \Gamma_{22}^1 f^\alpha_1 - \Gamma_{22}^2 f^\alpha_2 &= l_{22} N^\alpha. \end{aligned}$$

If we denote these component functions of f by u, v, w , then

$$\begin{aligned} N &= \frac{f_1 \times f_2}{\sqrt{\det(g_{ij})}} = \frac{(u_1, v_1, w_1) \times (u_2, v_2, w_2)}{\sqrt{\det(g_{ij})}} \\ &= \frac{(v_1 w_2 - v_2 w_1, w_1 u_2 - u_1 w_2, u_1 v_2 - u_2 v_1)}{\sqrt{\det(g_{ij})}}. \end{aligned}$$

So, for example, the third component N^3 of N satisfies

$$\begin{aligned} \det(g_{ij}) \cdot (N^3)^2 &= (u_1 v_2 - u_2 v_1)^2 \\ &= (u_1^2 + v_1^2)(u_2^2 + v_2^2) - (u_1 u_2 + v_1 v_2)^2 \\ &= (g_{11} - w_1^2)(g_{22} - w_2^2) - (g_{12} - w_1 w_2)^2 \\ &= \det(g_{ij}) - (g_{22} w_1^2 - 2g_{12} w_1 w_2 + g_{11} w_2^2). \end{aligned}$$

Using equations (1) for $\alpha = 3$, we obtain

$$\begin{aligned} (*) \quad &(w_{11} - \Gamma_{11}^1 w_1 - \Gamma_{11}^2 w_2)(w_{22} - \Gamma_{22}^1 w_1 - \Gamma_{22}^2 w_2) - (w_{12} - \Gamma_{12}^1 w_1 - \Gamma_{12}^2 w_2)^2 \\ &= (l_{11} l_{22} - l_{12}^2) \cdot \frac{\{\det(g_{ij}) - (g_{22} w_1^2 - 2g_{12} w_1 w_2 + g_{11} w_2^2)\}}{\det(g_{ij})} \\ &= K \{\det(g_{ij}) - (g_{22} w_1^2 - 2g_{12} w_1 w_2 + g_{11} w_2^2)\}, \end{aligned}$$

where the Γ 's and K are all computable in terms of the g_{ij} . We thus have a certain non-linear second order partial differential equation (*) for w . Notice that this equation does not contain w explicitly. If w is any solution, then so is $w + \text{constant}$, so we can always specify $w(0)$ arbitrarily.

It is easily checked (and is *a priori* clear on symmetry grounds) that u and v also satisfy equation (*). On the other hand, it is by no means true that (u, v, w) is a solution of our problem whenever u, v, w each satisfy (*), even if (u, v, w) is an immersion. In order to obtain more precise information, we must use a different procedure, due to Darboux. Suppose first that we are given an immersion $f = (u, v, w)$ such that I_f has components E, F, G ; thus

$$(1) \quad du \otimes du + dv \otimes dv + dw \otimes dw = E dx \otimes dx + F[dx \otimes dy + dy \otimes dx] + G dy \otimes dy,$$

where (x, y) is the standard coordinate system on \mathbb{R}^2 . By composing f with a Euclidean motion, if necessary, we can assume that

$$(2) \quad w_1(0, 0) = w_2(0, 0) = 0$$

$$\Downarrow$$

$$(3) \quad \begin{pmatrix} u_1 & u_2 \\ v_1 & v_2 \end{pmatrix} \text{ is nonsingular at } (0, 0).$$

Consider the tensor

$$\begin{aligned} \langle \cdot, \cdot \rangle' &= E dx \otimes dx + F[dx \otimes dy + dy \otimes dx] + G dy \otimes dy - dw \otimes dw \\ &= (E - w_1^2) dx \otimes dx + (F - w_1 w_2)[dx \otimes dy + dy \otimes dx] \\ &\quad + (G - w_2^2) dy \otimes dy. \end{aligned}$$

Using (1) we can write

$$(4) \quad \langle \cdot, \cdot \rangle' = du \otimes du + dv \otimes dv.$$

This is positive definite at $(0, 0)$ by (2), and hence positive definite in a neighborhood of $(0, 0)$. Moreover, (u, v) is a coordinate system for \mathbb{R}^2 in a neighborhood of $(0, 0)$, by (3). So equation (4) says that $\langle \cdot, \cdot \rangle'$ is flat, and thus has curvature $K' = 0$.

Recall (pg. II.131) that the metric with coefficients E, F, G has curvature K given by

$$(5) \quad K(EG - F^2)^2 = \det \begin{pmatrix} -\frac{1}{2}G_{11} + F_{12} - \frac{1}{2}E_{22} & \frac{1}{2}E_1 & F_1 - \frac{1}{2}E_2 \\ F_2 - \frac{1}{2}G_1 & E & F \\ \frac{1}{2}G_2 & F & G \end{pmatrix} \\ - \det \begin{pmatrix} 0 & \frac{1}{2}E_2 & \frac{1}{2}G_1 \\ \frac{1}{2}E_2 & E & F \\ \frac{1}{2}G_1 & F & G \end{pmatrix}.$$

To obtain the condition $K' = 0$, we set the right side equal to 0 after replacing E by $E - w_1^2$, etc. With the standard notation

$$\begin{aligned} p &= w_1, & q &= w_2, \\ r &= w_{11}, & s &= w_{12}, & t &= w_{22}, \end{aligned}$$

the (1, 1) term in the first matrix becomes

$$-\frac{1}{2}(G - w_2^2)_{11} + (F - w_1 w_2)_{12} - \frac{1}{2}(E - w_1^2)_{22} = -\frac{1}{2}G_{11} + F_{12} - \frac{1}{2}E_{22} + (s^2 - rt),$$

all third derivatives canceling. We thus obtain a second order equation for w , the “Darboux equation”, which written out explicitly becomes

$$\begin{aligned}
 (**) \quad 0 = & -4(EG - F^2)(rt - s^2) \\
 & + 2pr[2GF_2 - GG_1 - FG_2] + 2qr[EG_2 + FG_1 - 2FF_2] \\
 & + 4ps[FG_1 - GE_2] + 4qs[FE_2 - EG_1] \\
 & + 2pt[GE_1 + FE_2 - 2FF_1] + 2qt[2EF_1 - EE_2 - FE_1] \\
 & + (E - p^2)[E_2G_2 - 2F_1G_2 + (G_1)^2] \\
 & + (F - pq)[E_1G_2 - E_2G_1 - 2E_2F_2 - 2G_1F_1 + 4F_1F_2] \\
 & + (G - q^2)[G_1E_1 - 2F_2E_1 + (E_2)^2] \\
 & + 2[EG - F^2 - Gp^2 - Eq^2 + 2Fpq] \cdot [2F_{12} - E_{22} - G_{11}].
 \end{aligned}$$

Brute force computations will show that equations (*) and (**) are, in fact, the same (a somewhat more refined approach is given in Problem 4). But our derivation of (**) now enables us to relate solutions of (**) with functions $f = (u, v, w)$ satisfying (1). For suppose that w is a solution of (**) satisfying (2). Then $\langle \cdot, \cdot \rangle'$ is positive definite, and has curvature $K' = 0$. So there is a coordinate system (u, v) satisfying (4), which implies that (u, v, w) satisfies (1). The possible coordinate systems (u, v) for the flat metric $\langle \cdot, \cdot \rangle'$ all differ by a Euclidean motion of \mathbb{R}^2 , so (u, v) is determined by specifying

$$\begin{aligned}
 & u(0), \quad v(0), \\
 & u_1(0), \quad u_2(0), \quad v_1(0), \quad v_2(0),
 \end{aligned}$$

where the $u_i(0)$ and $v_i(0)$ have to be chosen so that (1) holds at $(0, 0)$. Notice that u and v will automatically satisfy (**), since this equation is equivalent to (*), which is satisfied by all component functions $f = (u, v, w)$ satisfying (1). We thus have the paradoxical situation that u, v, w all satisfy $(**) \equiv (*)$, but that once we pick the initial conditions w_1, w_2 along the x -axis which determine w , then we have almost no choice left for the initial conditions for u and v ; of course, we could just as well pick the initial conditions for u , say, and then be stuck with those for v and w .

The Darboux equation is not linear, but it is linear in $(rt - s^2)$, r, s, t ; as we mentioned in Chapter 10, section 8, equations of this sort are called “Monge-Ampère equations”. We can write our equation as

$$t(r + Ap + Bq) + C = 0,$$

where A, B, C do not involve t , and thus we can solve for t in terms of the other quantities,

$$(6) \quad t = \frac{-C}{r + Ap + Bq} = g(x, y, p, q, r, s).$$

More precisely, if we are given initial conditions along the x -axis such that $Ap + Bq + r \neq 0$ at $(0, 0)$, then we can write our equation in this form near $(0, 0)$. In Chapter 10, Part 4, we showed that the Cauchy problem for such an equation is equivalent to the Cauchy problem for a quasi-linear first order system, which can always be solved, by the Cauchy-Kowalewski theorem, if all functions in the equation, and the initial data, are analytic. Thus we see that the required isometric imbedding f exists locally if E, F, G are analytic.

Naturally we would like to know to what extent this restriction to analytic E, F, G and analytic initial data is necessary. Recall that a solution w of a second order PDE

$$F(x, y, w, p, q, r, s, t) = 0$$

is elliptic [respectively, hyperbolic] if and only if

$$4F_r F_t - F_s^2 > 0 \quad [\text{respectively, } < 0].$$

We consider the Darboux equation in the form (*) on page 143. Our condition becomes

$$4(w_{22} - \Gamma_{22}^1 w_1 - \Gamma_{22}^2 w_2)(w_{11} - \Gamma_{11}^1 w_1 - \Gamma_{11}^2 w_2) - 4(w_{12} - \Gamma_{12}^1 w_1 - \Gamma_{12}^2 w_2)^2 > 0 \quad [\text{respectively, } < 0].$$

Using equations (1) on page 143, this becomes

$$(N^3)^2(l_{22}l_{11} - l_{12}^2) > 0 \quad [\text{respectively, } < 0].$$

So at all points where $N^3 \neq 0$, the solution w is elliptic [hyperbolic] if and only if the corresponding surface (obtained by choosing u, v as before) has $K > 0$ [$K < 0$]. Thus the cases $K > 0$ and $K < 0$ require separate treatment.

We note first that Theorem 10-13 shows that if $K > 0$ everywhere and E, F, G are analytic, then every imbedding f such that I_f has components E, F, G is automatically analytic, so there is no point considering initial data which are not analytic. Expressed somewhat differently, if a surface $M \subset \mathbb{R}^3$ has $K > 0$ everywhere and a metric which is analytic in *some* coordinate system, then M is actually an analytic submanifold of \mathbb{R}^3 . In particular, if M is *any* (C^3) surface of constant positive curvature, then M is automatically analytic.

When E, F, G are not analytic, there is no known criterion on the initial data which will guarantee the existence of a solution of the Darboux equation (**). However, we might simply ask if there is *some* solution of the Darboux equation (without specifying initial conditions), and hence some imbedding f such that I_f has components E, F, G . The fact that a second order equation like (**) actually has solutions has been “known” for a long time—an actual proof may be found in Jacobowitz [1].

When $K < 0$ there is no problem. Theorem 10-12 shows that for any initial conditions with $F_t = r + Ap + Bq \neq 0$, there is always a solution w of (**) in a neighborhood of $(0, 0)$, and we can obtain solutions less differentiable than the functions E, F, G . (In the next chapter we will have occasion to examine the case where the initial conditions are such that $F_t = 0$.) We see, in particular, that there are surfaces of constant negative curvature in \mathbb{R}^3 which are C^∞ , but not analytic.

By the way, it is interesting to note that a surface of constant *mean* curvature H is always analytic (the sign of H couldn't be relevant, since it is not even well-determined). For suppose that $M \subset \mathbb{R}^3$ is a surface with $H = C$, given locally as the graph of a function h . Then formula (B') on pg. III.137 gives

$$\begin{aligned} 0 &= F(x, y, h, p, q, r, s, t) \\ &= (1 + q^2)r - 2pq s + (1 + p^2)t - 2C(1 + p^2 + q^2)^{3/2}, \end{aligned}$$

so

$$4F_r F_t - F_s^2 = 4[(1 + q^2)(1 + p^2) - p^2 q^2] = 4(1 + p^2 + q^2) > 0,$$

and h is analytic by Theorem 10-13.

We now consider the general problem of locally imbedding an n -manifold in \mathbb{R}^m . We are given g_{ij} on a neighborhood of $0 \in \mathbb{R}^n$, and we seek $f: U \rightarrow \mathbb{R}^m$, on some smaller neighborhood U , such that

$$\begin{aligned} (1) \quad g_{ij} &= \langle f_i, f_j \rangle \\ &= \sum_{\alpha=1}^m \langle f^\alpha_i, f^\alpha_j \rangle = \sum_{\alpha=1}^m \frac{\partial f^\alpha}{\partial x_i} \cdot \frac{\partial f^\alpha}{\partial x_j}. \end{aligned}$$

Since this is a set of $s_n = n(n+1)/2$ equations, it seems unlikely that we can always find f if $m < s_n$. In fact, if (1) is to hold, then all equations obtained from (1) by partial differentiation must also hold. If we evaluate these equations at 0, we obtain polynomial formulas expressing the derivatives

$$(2) \quad \frac{\partial^{r_1+\dots+r_n} g_{ij}}{\partial^{r_1} x_1 \dots \partial^{r_n} x_n} (0) \quad 0 \leq r_1 + \dots + r_n \leq r-1$$

in terms of the derivatives

$$(3) \quad \frac{\partial^{r_1+\dots+r_n} f^\alpha}{\partial^{r_1} x_1 \dots \partial^{r_n} x_n}(0) \quad 1 \leq r_1 + \dots + r_n \leq r.$$

For each g_{ij} , the number of derivatives in (2) is the binomial coefficient $\binom{n+r-1}{n}$ [= the number of ways of picking n things from $n+r-1$ things], for we can associate to each set $\alpha_1 < \alpha_2 < \dots < \alpha_n$ of integers from 1 to $n+r-1$ the numbers

$$r_1 = \alpha_1 - 1, \quad r_2 = \alpha_2 - \alpha_1 - 1, \quad \dots, \quad r_n = \alpha_n - \alpha_{n-1} - 1.$$

Thus,

$$\# \text{ of derivatives in (2) is } a = s_n \cdot \binom{n+r-1}{n}.$$

Similarly,

$$\# \text{ of derivatives in (3) is } b = m \cdot \left[\binom{n+r}{n} - 1 \right].$$

Now if $m < s_n$, then the first of these numbers will be greater than the second for large enough r . In fact,

$$(s_n - 1) \cdot \frac{(n+r)!}{n!r!} = s_n \cdot \frac{(n+r-1)!}{n!(r-1)!} \quad \text{for } r = n(s_n - 1).$$

But this means that the set of all possible derivatives (2), considered as a point in \mathbb{R}^a , is the image of a polynomial map defined on a lower dimensional space \mathbb{R}^b , so the derivatives (2) cannot be assigned arbitrarily for a map $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m < s_n$. In other words, not every g_{ij} can be obtained from some f .

It also seems reasonable to conjecture that we can always find an appropriate f when $m = s_n$. With the proper handling of subsidiary considerations, the proof of this conjecture can be reduced to the Cauchy-Kowalewski theorem (which means that we will have to assume that the g_{ij} are analytic). First a preliminary definition. Given $f: U \rightarrow \mathbb{R}^m$, consider the space spanned by the vectors

$$\frac{\partial f}{\partial x^i} \quad 1 \leq i \leq n, \quad \frac{\partial^2 f}{\partial x^i \partial x^j} \quad 1 \leq i, j \leq n$$

at a point $p \in U$ [this is the direct sum of the tangent space of $f(U)$ at $f(p)$ and the first normal space at $f(p)$, in the terminology of Addendum 4 to Chapter 7]. The map f is called **non-degenerate** if these $n + s_n$ vectors are linearly independent, for each $p \in U$. For example, a curve in $\mathbb{R}^2 \subset \mathbb{R}^m$ is non-degenerate if its curvature is nowhere zero.

9. THEOREM (BURSTIN-JANET-CARTAN). Let g_{ij} be the components of an analytic Riemannian metric in a neighborhood of $0 \in \mathbb{R}^n$. Then there is an analytic isometric imbedding $f: U \rightarrow \mathbb{R}^{s_n}$ (defined on some smaller neighborhood U).

PROOF. Let V_i be i -dimensional subspaces of \mathbb{R}^n_0 with

$$V_1 \subset \cdots \subset V_n = \mathbb{R}^n_0,$$

and let

$$H_i = \exp_0(V_i) \subset \mathbb{R}^n$$

(the exponential map being defined with respect to the metric given by the g_{ij}). Since H_1 is a curve, we can clearly find an analytic isometric imbedding $f_1: H_1 \rightarrow \mathbb{R}^{s_n}$; moreover, we can arrange for f_1 to be non-degenerate. We will now show that if $f_k: H_k \rightarrow \mathbb{R}^{s_n}$ is a free analytic isometric imbedding, then f_k can be extended to an analytic isometric imbedding $f_{k+1}: H_{k+1} \rightarrow \mathbb{R}^{s_n}$ (defined perhaps in a smaller neighborhood of 0). Moreover, for $k+1 < n$ we will show that f_{k+1} can be chosen to be non-degenerate [note that $l < n \implies l + s_l < s_n$]. This will clearly prove the theorem.

Step 1. By changing our coordinate system $x_1, \dots, x_k, y = x_{k+1}$ on H_{k+1} we can assume that

$$\begin{aligned} H_k &= \{(x, y) : y = 0\} \\ g_{i,k+1} &= 0 \quad (1 \leq i \leq k), \quad g_{k+1,k+1} = 1. \end{aligned}$$

Then the equations $g_{ij} = \langle f_i, f_j \rangle$ become

$$\begin{aligned} (1) \quad & \langle f_{x_i}, f_{x_j} \rangle = g_{ij} \\ & \langle f_{x_i}, f_y \rangle = 0 \\ & \langle f_y, f_y \rangle = 1. \end{aligned}$$

Differentiating the first equation with respect to y , and the second with respect to x_j , we find that if f satisfies (1), then it also satisfies the following set of equations, which are of first order with respect to the y variable:

$$(2a) \quad \langle f_y, f_{x_i x_j} \rangle = -\frac{1}{2}(g_{ij})_y$$

$$(2b) \quad \langle f_y, f_{x_i} \rangle = 0$$

$$(2c) \quad \langle f_y, f_y \rangle = 1.$$

Similar manipulations show that f also satisfies the equations

$$(3a) \quad \langle f_{yy}, f_{x_i} \rangle = 0$$

$$(3b) \quad \langle f_{yy}, f_y \rangle = 0$$

$$(3c) \quad \langle f_{yy}, f_{x_i x_j} \rangle = -\frac{1}{2}(g_{ij})_{yy} + \langle f_{yx_i}, f_{yx_j} \rangle,$$

which are of second order with respect to the y variables.

Conversely, suppose that f satisfies (3), and also satisfies (2) on H_k . We claim that f satisfies (1). First of all, since (3b) says that $\langle f_y, f_y \rangle_y = 0$, equation (2c) on H_k implies that $\langle f_y, f_y \rangle = 1$ everywhere. Consequently, we also have $\langle f_y, f_{x_i y} \rangle = 0$. So (3a) says that $\langle f_{x_i}, f_y \rangle_y = 0$, and then (2b) on H_k implies that $\langle f_{x_i}, f_y \rangle = 0$. Thus we have the last two equations of (1). Now from

$$\langle f_{x_i}, f_y \rangle = 0, \quad \langle f_{x_j}, f_y \rangle = 0, \quad \langle f_y, f_y \rangle = 1$$

we obtain

$$(i) \quad \langle f_{x_i x_j}, f_y \rangle + \langle f_{x_i}, f_{x_j y} \rangle = 0$$

$$(ii) \quad \langle f_{x_j x_i}, f_y \rangle + \langle f_{x_j}, f_{x_i y} \rangle = 0$$

$$(iii) \quad \langle f_{x_i y}, f_y \rangle = 0$$

and then

$$(iv) \quad \langle f_{x_i x_j y}, f_y \rangle + \langle f_{x_j x_i}, f_{yy} \rangle + \langle f_{x_j y}, f_{x_i y} \rangle + \langle f_{x_j}, f_{x_i y y} \rangle = 0 \quad \text{from (ii)}$$

$$(v) \quad \langle f_{x_i x_j y}, f_y \rangle + \langle f_{x_i y}, f_{x_j y} \rangle = 0 \quad \text{from (iii)}.$$

Equations (iv) and (v) give

$$\langle f_{x_i y y}, f_{x_j} \rangle = -\langle f_{x_i x_j}, f_{yy} \rangle,$$

so we have

$$\begin{aligned} \langle f_{x_i}, f_{x_j} \rangle_{yy} &= \langle f_{x_i y y}, f_{x_j} \rangle + 2\langle f_{x_i y}, f_{x_j y} \rangle + \langle f_{x_i}, f_{x_j y y} \rangle \\ &= -2\langle f_{x_i x_j}, f_{yy} \rangle + 2\langle f_{x_i y}, f_{x_j y} \rangle \\ &= (g_{ij})_{yy} \quad \text{by (3c).} \end{aligned}$$

On the other hand, (i) and (ii) give

$$\begin{aligned} \langle f_{x_i}, f_{x_j} \rangle_y &= -2\langle f_{x_i x_j}, f_y \rangle \\ &= (g_{ij})_y \quad \text{on } H_k, \quad \text{by (2a).} \end{aligned}$$

It follows that $\langle f_{x_i}, f_{x_j} \rangle_y = (g_{ij})_y$ everywhere. Since we have $\langle f_{x_i}, f_{x_j} \rangle = g_{ij}$ on H_k , we conclude that $\langle f_{x_i}, f_{x_j} \rangle = g_{ij}$ everywhere, as desired.

Step 2. Having established this, we now claim that there is an analytic function χ on H_k such that

$$(2') \quad \begin{aligned} \langle \chi, f_{x_i x_j} \rangle &= -\frac{1}{2}(g_{ij})_y \\ \langle \chi, f_{x_i} \rangle &= 0 \\ \langle \chi, \chi \rangle &= 1 \end{aligned} \quad \text{on } H_k,$$

χ is linearly independent of $f_{x_i}, f_{x_i x_j}$.

The reason for this is the following. At 0, we have $(g_{ij})_y = 0$ (Proposition II.4-1). So $\chi(0)$ is just a unit vector in \mathbb{R}^{s_n} which is perpendicular to all $f_{x_i}(0)$ and $f_{x_i x_j}(0)$ [such a vector exists, since $k + s_k < s_n$]. In general, we first pick a linear combination χ_1 of the (linearly independent) vectors $f_{x_i}, f_{x_i x_j}$ so that the first two conditions in (2') hold for χ_1 . Near 0, this makes χ_1 a vector of small norm. Then we add on an appropriate vector orthogonal to the f_{x_i} and $f_{x_i x_j}$ so that the norm becomes 1. There is no problem arranging for χ to be analytic.

Consider the following system of equations for functions $f, q: H_{k+1} \rightarrow \mathbb{R}^{s_n}$:

$$(*) \quad \begin{cases} f_y = q \\ \langle q_y, f_{x_i} \rangle = 0 & i = 1, \dots, k \\ \langle q_y, q \rangle = 0 \\ \langle q_y, f_{x_i x_j} \rangle = -\frac{1}{2}(g_{ij})_{yy} + \langle q_{x_i}, q_{x_j} \rangle & i, j = 1, \dots, k, \end{cases}$$

with the initial conditions

$$(*_0) \quad \begin{cases} f(x, 0) = f_k(x) \\ q(x, 0) = \chi(x) \end{cases} \quad \text{on } H_k.$$

If we have a solution (f, q) , then f will be a solution of (3) such that $\chi(x) = f_y(x, 0)$ satisfies (2) on H_k , so f will be an analytic isometric imbedding extending f_k in a neighborhood of 0. Now (*) is rather like the equations considered in the Cauchy-Kowalewski theorem, expressing the partials of the $2s_n$ functions $f^1, \dots, f^{s_n}, q^{s_1}, \dots, q^{s_n}$ with respect to y in terms of their partials with respect to x_1, \dots, x_k . However, we have more unknowns than equations (except for $k = n - 1$), and the q_y are not explicitly solved for. Such a problem is handled as follows.

Step 3. Write the last three sets of equations of (*) as a matrix equation

$$\begin{pmatrix} f_{x_i} \\ q \\ f_{x_i x_j} \end{pmatrix} \cdot \begin{pmatrix} q^1_y \\ \vdots \\ q^{s_n}_y \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ -\frac{1}{2}(g_{ij})_{yy} + \langle q_{x_i}, q_{x_j} \rangle \\ \vdots \end{pmatrix},$$

\uparrow
 $(k+1+s_k) \times s_n$
matrix

\uparrow
 $s_n \times 1$
matrix

\uparrow
 $(k+1+s_k) \times 1$
matrix

or for short as

$$B(f_{x_i}, q, f_{x_i x_j}) \cdot q_y = g.$$

On H_k , the rows of $B(f_{x_i}, q, f_{x_i x_j})$ are linearly independent, because f_k is non-degenerate, and by the choice (2') of χ . So this matrix has a right inverse. Moreover, we can pick this inverse analytically—that is, for any $r \leq s$ there is an analytic map $B \mapsto \tilde{B}$ from the $r \times s$ matrices of rank r to the $s \times r$ matrices such that

$$B \cdot \tilde{B} = r \times r \text{ identity matrix.}$$

[One specific way to define \tilde{B} is as follows. Write B as a collection of row vectors, $B = \begin{pmatrix} v_1 \\ \vdots \\ v_4 \end{pmatrix}$ for $v_i \in \mathbb{R}^s$. There is a unique decomposition $v_i = w_i + z_i$, where

$$w_i \in \text{subspace } W_i \subset \mathbb{R}^s \text{ spanned by } v_1, \dots, \hat{v}_i, \dots, v_r, \quad z_i \perp W_i;$$

clearly $z_i \neq 0$. Then

$$\langle v_i, z_j \rangle = \begin{cases} 0 & i \neq j \\ \langle z_i, z_i \rangle & i = j, \end{cases}$$

so we can choose \tilde{B} to be the matrix whose *columns* are $z_i / \langle z_i, z_i \rangle$.]

Now consider the system of equations

$$(**) \quad \begin{cases} f_y = q \\ q_y = \tilde{B}(f_{x_i}, q, f_{x_i x_j}) \cdot g, \end{cases}$$

with the initial conditions (*₀). This equation makes sense in a neighborhood

of H_k , since the rows of $B(f_{x_i}, q, f_{x_i x_j})$ are linearly independent there. Any solution of $(**)$ will be a solution of $(*)$, since

$$q_y = \tilde{B}(f_{x_i}, q, f_{x_i x_j}) \cdot g \implies B(f_{x_i}, q, f_{x_i x_j}) \cdot q_y = g.$$

But $(**)$ is a set of equations to which the Cauchy-Kowalewski theorem applies [more precisely, it can be reduced to such a set by the method of section 4 of Chapter 10]. Thus we have established the existence of the extension f .

Step 4. We still have to arrange for f to be non-degenerate when $k+1 < n$. We claim first that we can choose χ on H_k so that χ satisfies $(2')$, and also so that the vectors $\chi, \chi_{x_i}, f_{x_i}, f_{x_i x_j}$ are linearly independent at 0. To do this we again first choose χ_1 to be a linear combination of the $f_{x_i}, f_{x_i x_j}$ which satisfies the first two conditions of $(2')$; near 0 we have $|\chi_1| < 1$. We next choose $\alpha: H_k \rightarrow \mathbb{R}^{s_n}$ to be an analytic map with α perpendicular to $f_{x_i}, f_{x_i x_j}$ and $|\alpha|$ small. Then there is a constant λ_0 with $|\chi_1 + \lambda_0 \alpha| = 1$ at 0. We can assume, by renumbering, that the vectors of the set

$$A = \{\alpha, (\chi + \lambda_0 \alpha)_{x_1}, \dots, (\chi + \lambda_0 \alpha)_{x_h}, f_{x_i}, f_{x_i x_j}\}$$

are linearly independent at 0, and that $h \leq k$ is the largest integer with this property ($h = 0$ is a possibility, i.e., there may be no vectors $(\chi + \lambda_0 \alpha)_{x_i}$ in our set). If $h = k$ we are done. Otherwise, pick non-zero analytic functions $\beta_{h+1}, \dots, \beta_k$ which are orthogonal to the vectors of A , and also mutually orthogonal. Then determine the analytic function λ so that

$$|\chi(x)| = \left| \chi_1(x) + \lambda(x) \cdot \left(\alpha(x) + \sum_{i=h+1}^k \beta_i(x) \cdot x_i \right) \right| = 1.$$

Thus $\lambda_0 = \lambda(0)$. Suppose some linear combination of $\chi, \chi_{x_i}, f_{x_i}, f_{x_i x_j}$ vanishes at 0. I.e., suppose that at 0 we have

$$\begin{aligned} 0 = & a(\chi_1 + \lambda \alpha) + \sum_{i=1}^h a_i((\chi_1)_{x_i} + \lambda \alpha_{x_i} + \lambda_{x_i} \alpha) \\ & + \sum_{i=h+1}^k a_i((\chi_1)_{x_i} + \lambda \alpha_{x_i} + \lambda_{x_i} \alpha + \lambda \beta_i) \\ & + \sum_{i=1}^k b_i f_{x_i} + \sum_{i,j=1}^k b_{ij} f_{x_i x_j}. \end{aligned}$$

Take the inner product with some $\beta_i(0)$ ($i = h+1, \dots, k$). The β_i are mutually orthogonal, and β_i is perpendicular to all vectors of A ; moreover, each

$(\chi_1)_{x_i} + \lambda \alpha_{x_i}$ ($i = h+1, \dots, k$) is a linear combination of elements of A , by the maximality property of A . So we end up with

$$\lambda(0)a_i|\beta_i(0)|^2 = 0 \implies a_i = 0.$$

Then we also have $a = a_i = b_i = b_{ij} = 0$, since the vectors of A are linearly independent. Thus χ has all the required properties.

Now in a neighborhood of the $s_k + 2k + 1$ linearly independent vectors $\chi(0), \chi_{x_i}(0), f_{x_i}(0), f_{x_i x_j}(0)$ we can choose an analytic map $(v, v_i, w_i, w_{ij}) \mapsto h(v, v_i, w_i, w_{ij}) \neq 0$ such that

$$h(v, v_i, w_i, w_{ij}) \text{ is perpendicular to } v, v_i, w_i, w_{ij}.$$

Consider the equations

$$(**) \quad \begin{cases} f_y = q \\ q_y = \tilde{B}(f_{x_i}, q, f_{x_i x_j}) \cdot g + h(q, q_{x_i}, f_{x_i}, f_{x_i x_j}), \end{cases}$$

with the initial conditions $(*)_0$. A solution will again be a solution of $(*)$, since our choice of h gives

$$B(f_{x_i}, q, f_{x_i x_j}) \cdot h(q, q_{x_i}, f_{x_i}, f_{x_i x_j}) = 0.$$

The vectors $f_y, f_{yx_i}, f_{x_i}, f_{x_i x_j}$ are linearly independent near 0, by our choice of χ . Then $f_{yy} = q_y$ will be independent of $f_y, f_{yx_i}, f_{x_i}, f_{x_i x_j}$, since $h(f_y, f_{yx_i}, f_{x_i}, f_{x_i x_j})$ is linearly independent of all these vectors, while our explicit construction of \tilde{B} makes its columns span the same space as the rows of B , which implies that $\tilde{B}(f_{x_i}, q, f_{x_i x_j}) \cdot g$ is a linear combination of the $q, f_{x_i}, f_{x_i x_j}$. So f is non-degenerate. ♦

This proof can naturally be applied to the case $n = 2$, and then there is only the step from $k = 1$ to $k + 1 = 2$. In this case, the matrix $B(f_x, q, f_{xx})$ is a 3×3 invertible matrix, so we just consider the equations

$$\begin{aligned} f_y &= q \\ q_y &= B(f_x, q, f_{xx})^{-1} \cdot g. \end{aligned}$$

One can check that this is a hyperbolic system when $K < 0$, so that Theorem 10-12 can be applied, with the initial choice $f_1: (x\text{-axis}) \rightarrow \mathbb{R}^3$ being C^∞ rather than analytic; thus we can obtain the same results as we got by looking at the Darboux equation previously. Perhaps one could even try to analyze higher dimensional cases similarly, when the given metric has all sectional curvatures < 0 . There is not much interest in doing this, however, for although analyticity

was required to obtain the “best possible” local result of Theorem 9, there are global results where it is not needed. These results are essentially theorems in analysis, rather than geometry [with certain significant exceptions], and generally require rather involved techniques, some of which were created precisely for this problem. So we will merely indicate what these results are, and our discussion will be particularly brief since there are now several research reports which cover the field quite well.

One class of global results gives very strong information about the special case of surfaces in \mathbb{R}^3 . The first such question was raised by Hermann Weyl, who asked whether every metric $\langle \cdot, \cdot \rangle$ on S^2 with $K > 0$ comes from an isometric immersion in \mathbb{R}^3 (necessarily an imbedding as a convex surface, by Hadamard's Theorem). Although Weyl indicated an approach to this problem, the first affirmative solution, for analytic metrics, was given by H. Lewy [2]. A proof for C^k metrics, $k \geq 4$, was given by Nirenberg [1], and the cases $k = 2, 3$ were later handled by Heinz [1]. Already in 1942, A. D. Alexandrov had considered Weyl's problem from a completely different, totally geometric approach, involving polyhedral approximations to the surface. He was able to solve Weyl's problem for C^2 metrics, although his result did not indicate how differentiable the resulting surface would be when the metric was more differentiable. But this was established by later research, especially that of Pogorelov. At the same time, this pioneering work of Alexandrov led him to investigate arbitrary convex surfaces (which need not be smooth at all); although such surfaces may not have Riemannian metrics, we can still define an isometry between two such surfaces to be a homeomorphism preserving lengths of curves, and there are suitable generalizations of other differential geometric concepts like curvature (which may exist only almost everywhere). In consequence, there has developed an entirely disjoint school of differential geometry, whose practitioners are almost exclusively Russian mathematicians, which proves certain results in far greater generality than classical differential geometry, and has sometimes proved results from this field which are still inaccessible by the classical methods. Some examples of this will be mentioned in the next chapter, and the Bibliography gives further references to the Russian school.

In contrast to Weyl's problem, which arises by considering the metric $\langle \cdot, \cdot \rangle$ induced on S^2 by some imbedding of S^2 as a convex set in \mathbb{R}^3 , we now consider a strictly convex surface $M \subset \mathbb{R}^3$ and define a function $k > 0$ on S^2 by

$$K(p) = k(v(p)),$$

where $K > 0$ is the curvature of M , and the diffeomorphism $v: M \rightarrow S^2$ is

the normal map. This function k always satisfies certain integral equalities. To derive them, we note that we have

$$0 = \int_M v^i dA = \int_M \left\langle v, \frac{\partial}{\partial x^i} \right\rangle dA;$$

this follows from the Divergence Theorem (Problem I.9-13 or Theorem 7-57), applied to the region D bounded by M . Consequently, if da is the volume element of S^2 , and $x^i/k(x)$ denotes the function $x \mapsto x^i/k(x)$ on S^2 , then

$$\begin{aligned} (*) \quad \int_{S^2} \frac{1}{k(x)} \cdot x^i da &= \int_M v^* \left(\frac{1}{k(x)} \cdot x^i da \right) \\ &= \int_M \frac{1}{k \circ v} \cdot v^i \cdot v^*(da) \\ &= \int_M \frac{1}{K} \cdot v^i \cdot K dA = \int_M v^i dA \\ &= 0. \end{aligned}$$

“Minkowski’s problem” is to show that any function $k > 0$ on S^2 which satisfies the conditions (*) is $K \circ v^{-1}$ for some convex $M \subset \mathbb{R}^3$. This problem was solved by Lewy [3] in the analytic case, and by Nirenberg [1] in the C^2 case. It should also be mentioned that generalizations of Minkowski’s problem have been given in higher dimensions, in the style of the Russian school, by A. D. Alexandrov [3] and Fenchel and Jessen [1], but for the higher dimensional cases little is known about the differentiability of the hypersurfaces obtained.

Less delicate, but much more general, results are now available for the problem of isometrically imbedding arbitrary Riemannian manifolds in some Euclidean space. The first results along this line were by Nash [1], supplemented by Kuiper [1]. For a compact n -dimensional Riemannian manifold M , their results show that if M has any imbedding in \mathbb{R}^q , with $q \geq n+1$, then it also has a C^1 isometric imbedding. Thus compact orientable surfaces always have a C^1 isometric imbedding in \mathbb{R}^3 ; in particular, even the flat torus can be C^1 isometrically imbedded in \mathbb{R}^3 ! The most important isometric imbedding theorems stem from a second paper of Nash [2], where he proved that every C^∞ Riemannian manifold can be C^∞ isometrically imbedded in some Euclidean space. We will not give the dimensions of the Euclidean spaces involved; for this the reader may consult Gromov and Rokhlin [1], which gives a very complete discussion of the results known up to 1970. We merely mention that almost nothing is known about the lowest dimensional Euclidean space in which the imbedding is possible.

ADDENDUM

THE EMBEDDING PROBLEM VIA DIFFERENTIAL SYSTEMS

Although the general line of argument for the proof of Theorem 9 was proposed by Janet, it was Burstin who gave the first rigorous proof. The result is often known as the Cartan-Janet theorem because É. Cartan gave another (completely different) rigorous proof, using his theory of differential systems (Chapter 10, Addendum 1). We will give this proof here; so we assume that we have an analytic Riemannian metric in a neighborhood of $0 \in \mathbb{R}^n$, and we seek a local isometric imbedding into \mathbb{R}^{s_n} , $s_n = n(n+1)/2$.

Let $O(T\mathbb{R}^{s_n})$ be the bundle of orthonormal frames of \mathbb{R}^{s_n} , on which we have the dual forms ϕ^α and connection forms ψ_β^α ($1 \leq \alpha, \beta \leq s_n$); for simplicity we do not use bold-face letters for these forms on $O(T\mathbb{R}^{s_n})$. The forms $\phi^\alpha, \psi_\beta^\alpha$ give a basis for the dual space of the tangent space $O(T\mathbb{R}^{s_n})_u$ for any $u \in O(T\mathbb{R}^{s_n})$. Also let Z_1, \dots, Z_n be some fixed orthonormal moving frame on \mathbb{R}^n , with dual forms θ^i , connection forms ω_j^i , and curvature forms Ω_j^i . Suppose that $f: U \rightarrow \mathbb{R}^{s_n}$ is an isometry, for some neighborhood U of 0 in \mathbb{R}^n . Let $s = (Y_1, \dots, Y_n, Y_{n+1}, \dots, Y_{s_n})$ be any orthonormal moving frame on $f(U)$ with $Y_i = f_* Z_i$ for $i = 1, \dots, n$; then $s^* \phi^\alpha$ and $s^* \psi_\beta^\alpha$ are its dual forms and connection forms. Since f is an isometry, we clearly have

$$\begin{aligned}\theta^i &= f^*(s^* \phi^i) = (s \circ f)^* \phi^i & i &= 1, \dots, n \\ 0 &= f^*(s^* \phi^r) = (s \circ f)^* \phi^r & r &= n+1, \dots, s_n.\end{aligned}$$

Conversely, if $F: U \rightarrow O(T\mathbb{R}^{s_n})$ is a map which can be written as $F = s \circ f$, and

$$\begin{aligned}\theta^i &= F^* \phi^i & i &= 1, \dots, n \\ 0 &= F^* \phi^r & r &= n+1, \dots, s_n,\end{aligned}$$

then f is an isometry. We will look for F , and hence f , by looking for its graph in $\mathbb{R}^n \times O(T\mathbb{R}^{s_n})$. We have two projections

$$\begin{array}{ccc}\mathbb{R}^n \times O(T\mathbb{R}^{s_n}) & \xrightarrow{\pi_2} & O(T\mathbb{R}^{s_n}) \\ \downarrow \pi_1 & & \\ \mathbb{R}^n & & \end{array}$$

and for simplicity we will denote

$$\begin{aligned} \pi_1^* \theta^i & \text{ by } \theta^i, & \pi_1^* \omega_j^i & \text{ by } \omega_j^i, & \pi_1^* \Omega_j^i & \text{ by } \Omega_j^i, \\ \pi_2^* \phi^\alpha & \text{ by } \phi^\alpha, & \pi_2^* \psi_\beta^\alpha & \text{ by } \psi_\beta^\alpha. \end{aligned}$$

We easily see that our problem is solved if there is an n -dimensional manifold $\Gamma \subset \mathbb{R}^n \times O(T\mathbb{R}^{s_n})$ through some point $(0, u)$, such that $\pi_{1*}: \Gamma_{(0, u)} \rightarrow \mathbb{R}^n_0$ is one-one, and such that

$$\begin{aligned} \phi^i - \theta^i &= 0 & \text{ on } \Gamma & & i = 1, \dots, n \\ \phi^r &= 0 & \text{ on } \Gamma & & r = n+1, \dots, s_n. \end{aligned}$$

We want to find Γ as an integral manifold for an appropriate differential system \mathcal{L} . So, first of all, we want \mathcal{L} to contain the $\phi^i - \theta^i$ and the ϕ^r . Now

$$\begin{aligned} d(\phi^i - \theta^i) &= - \sum_{\alpha=1}^{s_n} \psi_\alpha^i \wedge \phi^\alpha + \sum_{j=1}^n \omega_j^i \wedge \theta^j \\ &= - \sum_{j=1}^n (\psi_j^i - \omega_j^i) \wedge \theta^j - \sum_{j=1}^n \psi_j^i \wedge (\phi^j - \theta^j) - \sum_{r=n+1}^{s_n} \psi_r^i \wedge \phi^r, \end{aligned}$$

and

$$\begin{aligned} d\phi^r &= - \sum_{\alpha=1}^{s_n} \psi_\alpha^r \wedge \phi^\alpha \\ &= - \sum_{j=1}^n \psi_j^r \wedge \theta^j - \sum_{j=1}^n \psi_j^r \wedge (\phi^j - \theta^j) - \sum_{t=n+1}^{s_n} \psi_t^r \wedge \theta^t, \end{aligned}$$

so in order to have $d\mathcal{L} \subset \mathcal{L}$ we also want the $\psi_j^i - \omega_j^i$ and the $\sum_j \psi_j^r \wedge \theta^j$ to be in \mathcal{L} . Similarly, since

$$\begin{aligned} d(\psi_j^i - \omega_j^i) &= - \sum_{\alpha=1}^{s_n} \psi_\alpha^i \wedge \psi_j^\alpha - \Omega_j^i + \sum_{h=1}^n \omega_h^i \wedge \omega_j^h \\ &= - \sum_{h=1}^n \psi_h^i \wedge (\psi_j^h - \omega_j^h) - \sum_{h=1}^n (\psi_h^i - \omega_h^i) \wedge \omega_j^h \\ &\quad - \sum_{r=n+1}^{s_n} \psi_r^i \wedge \psi_j^r - \Omega_j^i, \end{aligned}$$

we also want the $\sum_r \psi_r^i \wedge \psi_r^j - \Omega_j^i$ to be in \mathcal{L} . Moreover, we easily see that if \mathcal{L} is generated by

$$\begin{aligned}
 (a) \quad & \phi^i - \theta^i & i = 1, \dots, n \\
 (b) \quad & \phi^r & r = n+1, \dots, s_n \\
 (c) \quad & \psi_j^i - \omega_j^i & i, j = 1, \dots, n \\
 (d) \quad & \sum_{j=1}^n \psi_j^r \wedge \theta^j & r = n+1, \dots, s_n \\
 (e) \quad & \sum_{r=n+1}^{s_n} \psi_r^i \wedge \psi_r^j - \Omega_j^i & i, j = 1, \dots, n,
 \end{aligned}$$

then we have $d\mathcal{L} \subset \mathcal{L}$. The Cartan-Kähler Theorem (Theorem 10-15) tells us that the desired n -dimensional manifold $\Gamma \subset \mathbb{R}^n \times O(T\mathbb{R}^{s_n})$ exists if for some $u \in O(T\mathbb{R}^{s_n})$ there is an n -dimensional integral element $W \subset O(T\mathbb{R}^{s_n})_{(0,u)}$ of \mathcal{L} which contains a regular $(n-1)$ -dimensional integral element of \mathcal{L} , and for which $\pi_{1*}: W \rightarrow \mathbb{R}^n_0$ is one-one. We assume $n \geq 2$, since the case $n = 1$ is trivial.

We claim, first of all, that every point of $\mathbb{R}^n \times O(T\mathbb{R}^{s_n})$ is a regular 0-dimensional element of \mathcal{L} . To prove this we have to consider each $\mathcal{E}_1((x, u))$ for $(x, u) \in \mathbb{R}^n \times O(T\mathbb{R}^{s_n})$. By definition, $\mathcal{E}_1((x, u))$ is the set of all vectors (X, Y) [with $X \in \mathbb{R}^n_x$ and $Y \in O(T\mathbb{R}^{s_n})_u$] such that the forms (a)–(c) vanish on (X, Y) , that is:

$$(l) \quad \begin{cases} \phi^i(u)(Y) = \theta^i(x)(X) & i = 1, \dots, n \\ \phi^r(u)(Y) = 0 & r = n+1, \dots, s_n \\ \psi_j^i(u)(Y) = \omega_j^i(x)(X) & i, j = 1, \dots, n. \end{cases}$$

Because the $\phi^\alpha(u), \psi_\beta^\alpha(u)$ are a basis for $O(T\mathbb{R}^{s_n})_u$, the dimension of $\mathcal{E}_1((x, u))$ is always exactly the dimension of $\mathbb{R}^n \times O(T\mathbb{R}^{s_n})$ minus the number of forms (a)–(c), and thus a (non-zero) constant. So each (x, u) is a regular 0-dimensional integral element.

For any tangent vector $Y \in O(T\mathbb{R}^{s_n})_u$, it will be convenient to consider n vectors $Y^{(i)}$ in $\mathbb{R}^{s_n-n} = \mathbb{R}^{n(n-1)/2}$, defined by

$$Y^{(i)} = (\psi_i^{n+1}(u)(Y), \dots, \psi_i^{s_n}(u)(Y)).$$

Note that we can always choose $Y \in O(T\mathbb{R}^{s_n})_u$ satisfying (l) [i.e., with $(X, Y) \in \mathcal{E}_1((x, u))$] such that the $Y^{(i)}$ are any given n vectors in $\mathbb{R}^{n(n-1)/2}$.

Now at any point (x, u) , pick $(X_1, Y_1) \in \mathcal{E}_1((x, u))$ with X_1 a unit vector so that the $n - 1$ vectors

$$Y_1^{(1)}, \dots, Y_1^{(n-1)} \in \mathbb{R}^{n(n-1)/2}$$

are linearly independent [this is possible since $n(n - 1)/2 \geq n - 1$ for $n \geq 2$], and consider $\mathcal{E}_2((x, u), (X_1, Y_1))$. It is the set of all (X_2, Y_2) such that (l) holds, and such that the forms (d) and (e) vanish on the pair $(X_1, Y_1), (X_2, Y_2)$. If X_2 is a multiple of X_1 , then (l) implies that Y_2 is the same multiple of Y_1 , so we will assume that X_2 is linearly independent of X_1 . Since $(X_2, Y_2) \in \mathcal{E}_2((x, u), (X_1, Y_1))$ implies that any linear combination of (X_1, Y_1) and (X_2, Y_2) is also in $\mathcal{E}_2((x, u), (X_1, Y_1))$, in computing the dimension of this space we can restrict our attention to (X_2, Y_2) with X_1, X_2 orthonormal. Extend X_1, X_2 to an orthonormal basis X_1, \dots, X_n at u . Then (d) and (e) vanish on the pair $(X_1, Y_1), (X_2, Y_2)$ if and only if

$$(2) \quad Y_2^{(1)} = Y_1^{(2)}$$

$$(3) \quad Y_2^{(j)} \cdot Y_1^{(i)} - Y_2^{(i)} \cdot Y_1^{(j)} - \langle R(X_1, X_2)X_j, X_i \rangle = 0 \quad 1 \leq i < j \leq n,$$

where \cdot denotes the usual inner product in $\mathbb{R}^{n(n-1)/2}$. Equation (2) determines $Y_2^{(1)}$; then equation (3) for $i = 1, j = 2$ determines a hyperplane $H_2 \subset \mathbb{R}^{n(n-1)/2}$ in which $Y_2^{(2)}$ must lie; then equations (3) for $i = 1, j = 3$ and $i = 2, j = 3$ determine a plane $H_3 \subset \mathbb{R}^{n(n-1)/2}$ of codimension 2 in which $Y_2^{(3)}$ must lie; etc. [we use here the fact that $Y_1^{(1)}, \dots, Y_1^{(n-1)}$ are linearly independent]. In particular, the dimension of $\mathcal{E}_2((x, u), (X_1, Y_1))$ is the minimum possible. Thus (X_1, Y_1) generates a regular 1-dimensional integral element. Notice that $\mathcal{E}_2((x, u), (X_1, Y_1))$ does contain some (X_2, Y_2) [with X_1, X_2 orthonormal], since each H_α ($\alpha = 2, \dots, n$) has dimension

$$\frac{n(n-1)}{2} - (\alpha - 1) \geq \frac{n(n-1)}{2} - (n-1) \geq 0 \quad \text{for } n \geq 2.$$

In the case $n = 2$, we have just shown that there is some 2-dimensional integral element containing the regular 1-dimensional integral element generated by (X_1, Y_1) , which completes the proof. In the case $n \geq 3$ we claim that we can choose Y_2 so that

$$Y_1^{(1)}, \dots, Y_1^{(n-1)}, Y_2^{(2)}, \dots, Y_2^{(n-1)}$$

are linearly independent. We choose $Y_2^{(\alpha)}$ successively for $\alpha = 2, \dots, n-1$. We want $Y_2^{(\alpha)}$ to be linearly independent of the vectors in the set

$$A = \{Y_1^{(1)}, \dots, Y_1^{(n-1)}, Y_2^{(2)}, \dots, Y_2^{(\alpha-1)}\}.$$

Now $Y_2^{(\alpha)}$ must lie in the plane H_α , which is perpendicular to the vectors in the set

$$B = \{Y_1^{(1)}, \dots, Y_1^{(\alpha-1)}\}.$$

So we just need to have $\dim H_\alpha$ greater than the number of vectors in the set

$$A - B = \{Y_1^{(\alpha)}, \dots, Y_1^{(n-1)}, Y_2^{(2)}, \dots, Y_2^{(\alpha-1)}\}.$$

Thus we need

$$\dim H_\alpha = \frac{n(n-1)}{2} - (\alpha-1) > (n-\alpha) + (\alpha-2),$$

or

$$\frac{n(n-1)}{2} > (n-1) + (\alpha-2).$$

But for $\alpha \leq n-1$ we have

$$\begin{aligned} (n-1) + (\alpha-2) &\leq (n-1) + (n-3) \\ &< (n-1) + (n-2) \\ &\leq 1 + \dots + n-1 = \frac{n(n-1)}{2}. \end{aligned}$$

This proves the claim.

Now suppose that for some $k \leq n-1$ we have found $(X_1, Y_1), \dots, (X_k, Y_k)$ with X_1, \dots, X_k orthonormal such that

- (i) $(X_1, Y_1), \dots, (X_k, Y_k)$ generate a k -dimensional integral element,
- (ii) $(X_1, Y_1), \dots, (X_{k-1}, Y_{k-1})$ generate a regular $(k-1)$ -dimensional integral element,
- (iii) the $(n-1) + (n-2) + \dots + (n-k)$ vectors

$$Y_1^{(1)}, \dots, Y_1^{(n-1)}, Y_2^{(2)}, \dots, Y_2^{(n-1)}, \dots, Y_k^{(k)}, \dots, Y_k^{(n-1)}$$

are linearly independent.

[We have just done this for $k = 2$.] We claim that $(X_1, Y_1), \dots, (X_k, Y_k)$ generate a *regular* k -dimensional integral element, which can be extended to a $(k+1)$ -dimensional integral element generated by $(X_1, Y_1), \dots, (X_{k+1}, Y_{k+1})$ [with X_1, \dots, X_{k+1} orthonormal]; moreover, if $k+1 \leq n-1$, then Y_{k+1} can be picked so that $Y_{k+1}^{(k+1)}, \dots, Y_{k+1}^{(n-1)}$ are linearly independent of the vectors in (iii). Once we have proved this claim, it will clearly follow that there is

some n -dimensional integral element spanned by $(X_1, Y_1), \dots, (X_n, Y_n)$, with X_1, \dots, X_n orthonormal, such that $(X_1, Y_1), \dots, (X_{n-1}, Y_{n-1})$ span a regular $(n-1)$ -dimensional integral element. Thus the proof will be complete.

To calculate the dimension of $\mathcal{E}_{k+1}((x, u), (X_1, Y_1), \dots, (X_k, Y_k))$ we consider (X_{k+1}, Y_{k+1}) with X_1, \dots, X_{k+1} orthonormal, and again extend X_1, \dots, X_{k+1} to an orthonormal basis X_1, \dots, X_n at x . Then (d) and (e) vanish on the pairs $(X_h, Y_h), (X_{k+1}, Y_{k+1})$ ($1 \leq h \leq k$) if and only if

$$(2') \quad Y_{k+1}^{(h)} = Y_h^{(k+1)} \quad h = 1, \dots, k$$

$$(3') \quad Y_{k+1}^{(j)} \cdot Y_h^{(i)} - Y_{k+1}^{(i)} \cdot Y_h^{(j)} - \langle R(X_h, X_{k+1})X_j, X_i \rangle = 0 \quad \begin{array}{l} h = 1, \dots, k \\ 1 \leq i \leq j \leq n. \end{array}$$

Equations (2') determine $Y_{k+1}^{(h)}$ for $h \leq k$. We claim that with these values of $Y_{k+1}^{(h)}$, equation (3') holds for $i, j \leq k$. In fact, by hypothesis (i) we have

$$(a) \quad Y_h^{(i)} = Y_i^{(h)}, \quad Y_h^{(j)} = Y_j^{(h)} \quad i, j, h \leq k,$$

as well as

$$(b) \quad Y_i^{(l)} \cdot Y_j^{(\lambda)} - Y_i^{(\lambda)} \cdot Y_j^{(l)} - \langle R(X_j, X_i)X_l, X_\lambda \rangle = 0 \quad \begin{array}{l} i, j \leq k \\ \text{all } l, \lambda \leq n. \end{array}$$

Choose $l = h$ and $\lambda = k+1$ in (b), and substitute (2') and (a) into the equation. Using the identity $\langle R(X_j, X_i)X_h, X_{k+1} \rangle = \langle R(X_h, X_{k+1})X_j, X_i \rangle$, we obtain (3') for $i, j \leq k$.

Thus, we need to consider (3') only for i or $j \geq k+1$. Since we are choosing $i < j$, we have $j \geq k+1$ in either case. Moreover, we claim that for each $j \geq k+1$, and $h \leq k$, we need only consider the cases $h \leq i$. For if we have all these cases, and $i < h$, then by choosing i as our h , and h as our i , we have

$$(c) \quad Y_{k+1}^{(j)} \cdot Y_i^{(h)} - Y_{k+1}^{(h)} \cdot Y_i^{(j)} - \langle R(X_i, X_{k+1})X_j, X_h \rangle = 0.$$

Moreover, by (b) we also have

$$(d) \quad Y_h^{(k+1)} \cdot Y_i^{(j)} - Y_h^{(j)} \cdot Y_i^{(k+1)} - \langle R(X_i, X_h)X_{k+1}, X_j \rangle = 0.$$

In addition, (2') and (a) give

$$Y_h^{(k+1)} = Y_{k+1}^{(h)}, \quad Y_i^{(h)} = Y_h^{(i)}, \quad Y_i^{(k+1)} = Y_{k+1}^{(i)}.$$

So adding (c) and (d) gives

$$Y_{k+1}^{(j)} \cdot Y_h^{(i)} - Y_{k+1}^{(i)} \cdot Y_h^{(j)} - [\langle R(X_i, X_h)X_{k+1}, X_j \rangle + \langle R(X_i, X_{k+1})X_j, X_h \rangle] = 0.$$

Using the identities for the curvature tensor, we obtain finally

$$Y_{k+1}^{(j)} \cdot Y_h^{(i)} - Y_{k+1}^{(i)} \cdot Y_h^{(j)} - \langle R(X_h, X_{k+1})X_j, X_i \rangle = 0,$$

which is indeed just the required identity for i, j .

So consider now the equations

$$(3') \quad Y_{k+1}^{(j)} \cdot Y_h^{(i)} - Y_{k+1}^{(i)} \cdot Y_h^{(j)} - \langle R(X_h, X_{k+1})X_j, X_i \rangle = 0 \quad \begin{cases} h \leq k & i \geq h \\ i < j & j \geq k+1. \end{cases}$$

For $j = k+1$, there is one equation for each of the vectors

$$Y_1^{(1)}, \dots, Y_1^{(k)}, Y_2^{(2)}, \dots, Y_2^{(k)}, \dots, Y_k^{(k)},$$

which are linearly independent, by (iii). Thus $Y_{k+1}^{(k+1)}$ is restricted to lie in some plane $H_{k+1} \subset \mathbb{R}^{n(n-1)/2}$ of codimension $1 + \dots + k$. For $j = k+2$, there is then one equation for each of the linearly independent vectors

$$Y_1^{(1)}, \dots, Y_1^{(k+1)}, Y_2^{(2)}, \dots, Y_2^{(k+1)}, \dots, Y_k^{(k)}, Y_k^{(k+1)}.$$

So $Y_{k+1}^{(k+2)}$ is restricted to lie in some plane H_{k+2} of codimension $2 + \dots + (k+1)$. Etc. We see right away that

$$\dim \mathcal{E}_{k+1}((x, u), (X_1, Y_1), \dots, (X_k, Y_k))$$

is the minimum possible, so $(X_1, Y_1), \dots, (X_k, Y_k)$ do generate a regular k -dimensional integral element. Moreover, it can be extended to a $(k+1)$ -dimensional integral element, by choosing an appropriate (X_{k+1}, Y_{k+1}) [with X_1, \dots, X_{k+1} orthonormal], since each H_α ($\alpha = k+1, \dots, n$) has dimension

$$\begin{aligned} & \frac{n(n-1)}{2} - [(\alpha - k) + \dots + (\alpha - 1)] \\ &= 1 + \dots + (n-1) - [(\alpha - k) + \dots + (\alpha - 1)] \\ &\geq 0. \end{aligned}$$

We claim, finally, that if $k+1 \leq n-1$, then Y_{k+1} can be picked so that $Y_{k+1}^{(k+1)}, \dots, Y_{k+1}^{(n-1)}$ are linearly independent of the vectors in (iii). We pick $Y_{k+1}^{(\alpha)}$ successively, for $\alpha = k+1, \dots, n-1$. The vector $Y_{k+1}^{(\alpha)}$ has to be picked linearly independent of the vectors in the set

$$A = \{Y_1^{(1)}, \dots, Y_1^{(n-1)}, \dots, Y_k^{(k)}, \dots, Y_k^{(n-1)}, Y_{k+1}^{(k+1)}, \dots, Y_{k+1}^{(\alpha-1)}\},$$

with cardinality $(n-k) + \dots + (n-1) + (\alpha - k - 1)$.

Equations (3') say that $Y_{k+1}^{(\alpha)}$ must lie in a plane H_α perpendicular to the vectors of the set

$$B = \{Y_1^{(1)}, \dots, Y_1^{(\alpha-1)}, Y_2^{(2)}, \dots, Y_2^{(\alpha-1)}, \dots, Y_k^{(k)}, \dots, Y_k^{(\alpha-1)}\},$$

with cardinality r , say.

This is possible if $\dim H_\alpha$ is greater than the number of vectors in the difference set $A - B$. Since r is just the codimension of H_α , we thus need to have

$$\frac{n(n-1)}{2} - r > (n-k) + \dots + (n-1) + (\alpha - k - 1) - r.$$

But for $\alpha \leq n-1$ we have

$$\begin{aligned} (n-k) + \dots + (n-1) + (\alpha - k - 1) &\leq (n-k) + \dots + (n-1) + (n-k-2) \\ &< (n-k) + \dots + (n-1) + (n-k-1) \\ &\leq 1 + \dots + (n-1) = \frac{n(n-1)}{2}, \end{aligned}$$

as required.

PROBLEMS

1. Let $a_i \neq 0$ for $1 < i < n - 1$ with $\sum_i a_i^2 = 1$. Define an immersion

$$f: \{x \in \mathbb{R}^n : x_n < 0\} \rightarrow \mathbb{R}^{2n-1}$$

by

$$\begin{aligned} f^{2i-1}(x) &= a_i e^{x_n} \cos(x_i/a_i) \\ f^{2i}(x) &= a_i e^{x_n} \sin(x_i/a_i) & 1 \leq i \leq n-1 \\ f^{2n-1}(x) &= \int_0^{x_n} \sqrt{1 - e^{2t}} dt. \end{aligned}$$

Calculate that the induced metric has constant negative curvature.

2. Let M^n be a manifold of constant curvature K isometrically immersed in a manifold N^{2n-1} of constant curvature $K_0 > K$.

(a) Using equation (1) on page 141, generalize the argument in the second proof of Lemma 5-10 to prove that the bracket of two unit asymptotic vector fields on M is zero.

(b) Also use an argument from this proof to show that if M is complete, then its universal covering space must be \mathbb{R}^n .

(c) Conclude that we cannot have $K > 0$.

3. (a) Let $T: V \rightarrow V$ be a self-adjoint linear transformation, and let $V = V_1 \oplus \cdots \oplus V_k$ where the V_i are the mutually orthogonal eigenspaces for the distinct eigenvalues $\lambda_1, \dots, \lambda_k$. Let $P_i: V \rightarrow V_i$ be the corresponding orthogonal projections, $P_i(\sum_j a_j v_j) = a_i v_i$. Show that P_i is a polynomial in T , namely,

$$\frac{(T - \lambda_1) \cdots (T - \lambda_{i-1})(T - \lambda_{i+1}) \cdots (T - \lambda_k)}{(\lambda_i - \lambda_1) \cdots (\lambda_i - \lambda_{i-1})(\lambda_i - \lambda_{i+1}) \cdots (\lambda_i - \lambda_k)}.$$

Thus we have $T = \sum_i \lambda_i P_i$ where the P_i are polynomials in T .

(b) Let $S: V \rightarrow V$ be another self-adjoint linear transformation with $S = \sum_j \mu_j Q_j$ for polynomials Q_j in S . If S and T commute, then all P_i and Q_j commute. Let $A = \sum_{ij} a_{ij} P_i Q_j$ for distinct $a_{ij} \in \mathbb{R}$. Show that A is a self-adjoint transformation, that S and T are both polynomials in A , and that any linear transformation that commutes with S and T also commutes with A .

(c) If T_1, \dots, T_r are commuting self-adjoint operators, then there is a self-adjoint transformation A such that each T_i is a polynomial in A .

(d) Consequently, T_1, \dots, T_r can be simultaneously diagonalized.

4. (a) Let $f: M \rightarrow \mathbb{R}$, where $(M, \langle \cdot, \cdot \rangle)$ is a Riemannian manifold, and consider the symmetric covariant tensor $\nabla(df)$ of order 2, with components $f_{i;j}$ in a coordinate system. Each $\nabla(df)(p): M_p \times M_p \rightarrow \mathbb{R}$ can be regarded as a linear transformation $M_p \rightarrow M_p$, by using the inner product on M_p . So we can form $\mathcal{D}f(p) = \det(\nabla(df)(p))$. Equivalently, $\mathcal{D}f(p) = \det(\nabla(df)(p)(X_i, X_j))$, where X_1, \dots, X_n is any orthonormal basis of M_p . In a coordinate system we have

$$\mathcal{D}f = \frac{\det(f_{i;j})}{\det(g_{ij})}.$$

Show that equation (*) on page 143 can be written

$$\mathcal{D}w = K(1 - \Delta_1 w).$$

(b) Check that equations (*) and (**) are the same when $F = 0$, and conclude that they are always the same.

CHAPTER 12

RIGIDITY

In Chapter 7 we proved a result (Theorem 7-47) which is but a special case of the following more general

1. THEOREM. Let M and \bar{M} be immersed hypersurfaces in \mathbb{R}^{n+1} , and let $\phi: M \rightarrow \bar{M}$ be an isometry. Suppose that $d\nu: M_p \rightarrow M_p$ has rank ≥ 3 . Then $(\phi^*\bar{\Pi})(p) = \pm\Pi(p)$. [This equation makes sense even though ν and $\bar{\nu}$ may be defined only locally, and then only up to sign.]

Consequently, if M and \bar{M} are connected (not necessarily complete) hypersurfaces and $d\nu: M_p \rightarrow M_p$ has rank ≥ 3 for all $p \in M$, then ϕ is the restriction of a Euclidean motion.

PROOF. To deduce the second part of the theorem from the first part, we recall (pg. IV.63) that there is an inner product preserving bundle isomorphism $\tilde{\phi}: \text{Nor } M \rightarrow \text{Nor } \bar{M}$ covering ϕ . The first part of the theorem shows that if $p \in M$, then

$$\begin{array}{lll} \text{either} & \bar{s}(\phi_*X, \phi_*Y) = \tilde{\phi}(s(X, Y)) & \text{for all } X, Y \in M_p \\ \text{or} & \bar{s}(\phi_*X, \phi_*Y) = -\tilde{\phi}(s(X, Y)) & \text{for all } X, Y \in M_p. \end{array}$$

Moreover, only one alternative can hold at each p , since $d\nu: M_p \rightarrow M_p$ is non-singular. It follows that one of the alternatives holds for all $p \in M$. Then Theorem 7-21 shows that f is the restriction of a Euclidean motion.

To prove the first part, let $X_1, \dots, X_n \in M_p$ be an orthonormal basis, and define the $n \times n$ symmetric matrix S by $S_{ij} = \Pi(X_i, X_j)$. Similarly, let $\bar{X}_i = \phi_*X_i \in \bar{M}_{f(p)}$ and define the symmetric $n \times n$ matrix \bar{S} by $\bar{S}_{ij} = \bar{\Pi}(\bar{X}_i, \bar{X}_j)$. Gauss' equation shows that

$$\begin{aligned} S_{i_1j_1}S_{i_2j_2} - S_{i_1j_2}S_{i_2j_1} &= \langle R(X_{i_1}, X_{i_2})X_{j_2}, X_{j_1} \rangle \\ &= \langle \bar{R}(\bar{X}_{i_1}, \bar{X}_{i_2})\bar{X}_{j_2}, \bar{X}_{j_1} \rangle, & \text{since } \phi \text{ is an isometry} \\ &= \bar{S}_{i_1j_1}\bar{S}_{i_2j_2} - \bar{S}_{i_1j_2}\bar{S}_{i_2j_1}. \end{aligned}$$

We now use an algebraic

2. LEMMA. Let S and \bar{S} be symmetric $n \times n$ matrices, with $\text{rank } S \geq 3$. Suppose that the determinant of every 2×2 submatrix of S equals the determinant of the corresponding 2×2 submatrix of \bar{S} . Then $\bar{S} = \pm S$.

PROOF. To isolate the main idea of the proof, we first consider

Case 1. The matrices S and \bar{S} are non-singular. (Then the hypothesis on the rank just means that $n \geq 3$.) Let $T, \bar{T}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the non-singular linear transformations with matrices S, \bar{S} . Then $T^*, \bar{T}^*: \mathbb{R}^{n*} \rightarrow \mathbb{R}^{n*}$ are also non-singular. We can also consider the linear transformations

$$T^*, \bar{T}^*: \Omega^2(\mathbb{R}^n) \rightarrow \Omega^2(\mathbb{R}^n).$$

If $\phi_1, \dots, \phi_n \in \mathbb{R}^{n*}$ is the dual basis to the standard basis of \mathbb{R}^n , then

$$\begin{aligned} T^*(\phi_{i_1} \wedge \phi_{i_2}) &= T^*\phi_{i_1} \wedge T^*\phi_{i_2} \\ &= \sum_{j_1} S_{i_1 j_1} \phi_{j_1} \wedge \sum_{j_2} S_{i_2 j_2} \phi_{j_2} \\ &= \sum_{j_1 < j_2} (S_{i_1 j_1} S_{i_2 j_2} - S_{i_1 j_2} S_{i_2 j_1}) \cdot \phi_{j_1} \wedge \phi_{j_2}, \end{aligned}$$

and similarly for \bar{T}^* . Thus we see that the hypotheses on the determinants of S and \bar{S} is equivalent to the assertion that

$$(1) \quad T^* = \bar{T}^*: \Omega^2(\mathbb{R}^n) \rightarrow \Omega^2(\mathbb{R}^n).$$

Now given any $\phi \in \mathbb{R}^{n*}$, we claim that $T^*\phi$ and $\bar{T}^*\phi$ must be linearly dependent. For otherwise we could choose $\psi \in \mathbb{R}^{n*}$ with $T^*\phi, \bar{T}^*\phi, T^*\psi$ linearly independent. Then we would have

$$\begin{aligned} 0 \neq T^*\phi \wedge T^*\psi \wedge \bar{T}^*\phi &= T^*(\phi \wedge \psi) \wedge \bar{T}^*\phi \\ &= \bar{T}^*(\phi \wedge \psi) \wedge \bar{T}^*\phi \quad \text{by (1)} \\ &= \bar{T}^*\phi \wedge \bar{T}^*\psi \wedge \bar{T}^*\phi \\ &= 0, \end{aligned}$$

a contradiction. Thus $\bar{T}^*\phi = c \cdot T^*\phi$ for some $c \in \mathbb{R}$. If we choose linearly independent $\phi_1, \phi_2 \in \mathbb{R}^{n*}$, and apply this result to $\phi_1, \phi_2, \phi_1 + \phi_2$, we find that there are constants c_1, c_2, c with

$$\begin{aligned} \bar{T}^*\phi_1 &= c_1 \cdot T^*\phi_1, & \bar{T}^*\phi_2 &= c_2 \cdot T^*\phi_2 \\ \bar{T}^*\phi_1 + \bar{T}^*\phi_2 &= c \cdot (T^*\phi_1 + T^*\phi_2). \end{aligned}$$

It follows that $c_1 = c_2$. So $\bar{T}^* = c \cdot T^*$ for some $c \in \mathbb{R}$. From (1) we see that $c = \pm 1$.

Case 2. General Case. Since S is symmetric, the map $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is self-adjoint with respect to the usual metric on \mathbb{R}^n ; that is, $\langle Tv, w \rangle = \langle v, Tw \rangle$ for $v, w \in \mathbb{R}^n$. Similarly, if we give \mathbb{R}^{n*} the inner product with $\langle \phi_i, \phi_j \rangle = \delta_{ij}$, then

$$(2) \quad \langle T^*\phi, \psi \rangle = \langle \phi, T^*\psi \rangle \quad \text{for } \phi, \psi \in \mathbb{R}^{n*}.$$

Let $(\ker T^*)^\perp \subset \mathbb{R}^{n*}$ be the orthogonal complement of $\ker T^* \subset \mathbb{R}^{n*}$ with respect to this inner product. We easily see from (2) that T^* takes $(\ker T^*)^\perp$ into itself and that

$$(3) \quad T^*: (\ker T^*)^\perp \rightarrow (\ker T^*)^\perp \quad \text{is one-one.}$$

We now claim:

$$(4) \quad \ker T^* = \{\phi : T^*\phi \wedge T^*\psi = 0 \text{ for all } \psi \in \mathbb{R}^{n*}\} = W, \quad \text{say.}$$

It is clear that $\ker T^* \subset W$. Conversely, given $\phi \in W$, write

$$\phi = \phi_1 + \phi_2 \quad \text{with} \quad \phi_1 \in \ker T^* \subset W, \quad \phi_2 \in (\ker T^*)^\perp.$$

Then

$$(5) \quad \phi_2 \in (\ker T^*)^\perp \cap W.$$

We want to show that $\phi_2 = 0$. Note that

$$\dim(\ker T^*)^\perp = \text{rank } T^* = \text{rank } S \geq 3,$$

so if $\phi_2 \neq 0$, then there is $\psi \in (\ker T^*)^\perp$ with ϕ_2 and ψ linearly independent. By (3), this means that $T^*\phi_2$ and $T^*\psi$ are linearly independent, so

$$0 \neq T^*\phi_2 \wedge T^*\psi,$$

contradicting the fact that $\phi_2 \in W$ [by (5)]. Thus ϕ_2 must be 0, and we have demonstrated (4).

Notice that in proving (4) we really used only the fact that $\text{rank } S \geq 2$. Now we also have $\text{rank } \bar{S} \geq 2$ (otherwise every 2×2 submatrix of \bar{S} would have determinant 0). So we also have

$$(4) \quad \ker \bar{T}^* = \{\phi : \bar{T}^*\phi \wedge \bar{T}^*\psi = 0 \text{ for all } \psi \in \mathbb{R}^{n*}\}.$$

Then equation (1) shows that $\ker T^* = \ker \bar{T}^*$. Now we can apply Case 1 to

$$T^*, \bar{T}^*: (\ker T^*)^\perp \rightarrow (\ker T^*)^\perp,$$

for these maps are one-one by (3) and the corresponding (3), and moreover $\dim(\ker T^*)^\perp = \text{rank } T^* = \text{rank } S \geq 3$. ♦

The rank of $d\nu: M_p \rightarrow M_p$ is called the **type number** $t(p)$ of M at p ; it is the number of non-zero principal curvatures at p . The hypothesis that $t(p) \geq 3$ says, roughly speaking, that M curves in at least 3 different directions at p . The hypersurfaces with $t(p) = 0$ or 1 for all p are precisely the flat hypersurfaces, with curvature tensor $R = 0$, while the hypersurfaces with $t(p) = 2$ for all p may be regarded as a sort of generalization of this class. They have been classified into three different types by É. Cartan [1], but the classification suffers the same defect as the classical classification of flat surfaces, for there is no discussion of the manner in which hypersurfaces of different types can be joined together. We will have a little more to say about this later on.

Theorem 1 is often expressed by saying that a hypersurface in \mathbb{R}^{n+1} which bends enough is “rigid”. The first precise proof was by Killing [1], although the result had been stated by Beez [1], who found it so astounding that he could barely cease discussing it, and practically regarded it as a proof that space can’t be 4-dimensional. While we might not be willing to go quite so far as that, it is nevertheless true that because of this result most of the interest in rigidity phenomena has centered on the case of surfaces in \mathbb{R}^3 , where intuition tells us that a small piece of surface is not rigid, and at the same time suggests that compact surfaces should be rigid. Actually, there are several different senses in which a surface can be rigid. Books written in English often consider only one possible sense, or tend to be rather sloppy about distinguishing the various possibilities. I therefore propose to introduce some terminology which, although it may not be especially appealing aesthetically, and suffers the disadvantage of not being standard, at least has the virtue of being unambiguous.

Consider a C^∞ imbedding $f: M \rightarrow \mathbb{R}^3$. The strongest sense in which $f(M)$ can be “bent” corresponds to the ordinary conception of the word, whereby $f(M)$ passes continuously from one shape to another, without being stretched. To express this idea precisely, we define a **bending** of the imbedding $f: M \rightarrow \mathbb{R}^3$ to be a C^∞ map $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ such that

- (a) each map $\bar{\alpha}(t): M \rightarrow \mathbb{R}^3$, given by $p \mapsto \alpha(t, p)$, is an imbedding,
- (b) $\bar{\alpha}(0) = f$,
- (c) $\bar{\alpha}(t)^*(\langle \cdot, \cdot \rangle) = \bar{\alpha}(0)^*(\langle \cdot, \cdot \rangle)$ for all $t \in [0, 1]$.

Thus α is a “variation” of f , in the terminology of Chapter 9. To be a little more precise, α should be called a *bending through imbeddings*, and we can also define a bending through immersions. The bending $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ is called **trivial** if each $\bar{\alpha}(t)$ is $A_t \circ f$ for some Euclidean motion A_t ; it is called **non-trivial** if at least one $\bar{\alpha}(t)$ is not of this form. We say that the imbedding $f: M \rightarrow \mathbb{R}^3$ is **bendable** if there is a non-trivial bending of f ; otherwise it is called **unbendable**. To be

precise, we must speak of bendability and unbendability through imbeddings or through immersions. We can also define when an immersion $f: M \rightarrow \mathbb{R}^3$ is bendable or unbendable; in this case, of course, only bendings through immersions can be relevant. (It is also possible to consider C^l bendings of C^k imbeddings and immersions, for $1 \leq l \leq k \leq \omega$; but we shall hardly ever stray from the case $k = l = \infty$.) Finally, a submanifold $M \subset \mathbb{R}^3$ is called bendable or unbendable (through imbeddings or immersions) according as whether the inclusion map $i: M \rightarrow \mathbb{R}^3$ is bendable or unbendable.

One way of modifying the concept of a bending is by taking a discrete analogue: We will call an imbedding $f: M \rightarrow \mathbb{R}^3$ **warpable** if there is an imbedding $g: M \rightarrow \mathbb{R}^3$ such that $f^*\langle \cdot, \cdot \rangle = g^*\langle \cdot, \cdot \rangle$, but such that g is not $A \circ f$ for any Euclidean motion A . If f is not warpable, it will be called **unwarpable**. We can also define a warpable immersion. It is conceivable that there is an imbedding $f: M \rightarrow \mathbb{R}^3$ such that

- (i) there exists an *immersion* $g: M \rightarrow \mathbb{R}^3$ with $f^*\langle \cdot, \cdot \rangle = g^*\langle \cdot, \cdot \rangle$,
- (ii) there does not exist an *imbedding* $g: M \rightarrow \mathbb{R}^3$ with $f^*\langle \cdot, \cdot \rangle = g^*\langle \cdot, \cdot \rangle$ except for g of the form $A \circ f$ for some Euclidean motion A ;

we can express this by saying that f is warpable as an immersion, but not as an imbedding (an actual example of this phenomenon will be mentioned later). A surface $M \subset \mathbb{R}^3$ is called warpable or unwarpable (as an imbedding or immersion) according as whether the inclusion map $i: M \rightarrow \mathbb{R}^3$ is warpable or unwarpable. An unwarpable surface $M \subset \mathbb{R}^3$ is sometimes called “uniquely determined” (for M is then uniquely determined, up to a Euclidean motion, by its induced metric); similarly, we can speak of an imbedding or immersion $f: M \rightarrow \mathbb{R}^3$ being “uniquely determined”. A bendable surface is obviously warpable, but it is not *a priori* clear whether there are any warpable surfaces which are not bendable.*

We can also consider an infinitesimal analogue of a bending α , by looking at its “variation vector field” Z . This is the vector field along f defined by

$$Z(p) = \text{tangent vector at } 0 \text{ of } t \mapsto \alpha(t, p) \in \mathbb{R}^3_{f(p)}.$$

*As if matters were not already sufficiently complicated, one more possibility must be mentioned, which for the sake of simplicity we shall describe in terms of submanifolds, rather than imbeddings. Let $M \subset \mathbb{R}^3$ be a surface. It is conceivable that M is warpable, so that there is an isometry $\phi: M \rightarrow \bar{M} \subset \mathbb{R}^3$ which is not the restriction of a Euclidean motion, but that whenever we have an isometry $\phi: M \rightarrow \bar{M}$ then there is also another isometry $\psi: M \rightarrow \bar{M}$ which is the restriction of a Euclidean motion. Thus M might be warpable, but only into surfaces which happen to be congruent to M . No example of such a phenomenon is known, however.

Consider for the moment the case where $M \subset \mathbb{R}^3$ and $f = \text{inclusion map}$. Since α satisfies

$$\langle X, Y \rangle = \langle \bar{\alpha}(t)_*(X), \bar{\alpha}(t)_*(Y) \rangle$$

for all t , and $X, Y \in M_p$, we have

$$\begin{aligned} (1) \quad 0 &= \frac{d}{dt} \langle \bar{\alpha}(t)_*(X), \bar{\alpha}(t)_*(Y) \rangle \\ &= \left\langle \frac{D'}{\partial t} \bar{\alpha}(t)_*(X), \bar{\alpha}(t)_*(Y) \right\rangle + \left\langle \bar{\alpha}(t)_*(X), \frac{D'}{\partial t} \bar{\alpha}(t)_*(Y) \right\rangle, \end{aligned}$$

where $D'/\partial t$ denotes covariant differentiation in the ambient space \mathbb{R}^3 , as usual. Now if c is a curve in M with $c'(0) = X$, then

$$\begin{aligned} \left. \frac{D'}{\partial t} \right|_{t=0} \bar{\alpha}(t)_*(X) &= \left. \frac{D'}{\partial t} \right|_{t=0} \left. \frac{\partial}{\partial s} \right|_{s=0} \alpha(t, c(s)) \\ &= \left. \frac{D'}{\partial s} \right|_{s=0} \left. \frac{\partial}{\partial t} \right|_{t=0} \alpha(t, c(s)) && \text{by Proposition II.6-9} \\ &&& \text{(or simply equality of mixed partials)} \\ &= \left. \frac{D'}{\partial s} \right|_{s=0} Z(c(s)) \\ &= \nabla'_X Z. \end{aligned}$$

So equation (1) becomes

$$(2) \quad 0 = \langle \nabla'_X Z, Y \rangle + \langle X, \nabla'_Y Z \rangle \quad \text{for all } X, Y \text{ tangent to } M.$$

This is equivalent, by polarization, to

$$(2') \quad 0 = \langle \nabla'_X Z, X \rangle \quad \text{for all } X \text{ tangent to } M,$$

and these equations can also be written

$$(2'') \quad \begin{cases} 0 = \langle dZ(X), X \rangle & \text{for all } X \text{ tangent to } M \\ 0 = \langle dZ(X), Y \rangle + \langle X, dZ(Y) \rangle & \text{for all } X, Y \text{ tangent to } M, \end{cases}$$

where Z is considered as an \mathbb{R}^3 -valued function on M , and the tangent vector X of \mathbb{R}^3 is identified with an element of \mathbb{R}^3 .

For the general case of an immersion $f: M \rightarrow \mathbb{R}^3$, the vector field Z along f still satisfies (2), except that now the term

$$\nabla'_X Z = \left. \frac{D'}{\partial s} \right|_{s=0} Z(c(s)) = \left(\left. \frac{dZ^1(c(s))}{ds} \right|_{s=0}, \dots, \left. \frac{dZ^3(c(s))}{ds} \right|_{s=0} \right)$$

denotes a “covariant derivative of a vector field along f ”. Equation (2'') becomes

$$(2''') \quad \begin{aligned} 0 &= \langle dZ(X), f_*(X) \rangle \\ &= \langle dZ(X), df(X) \rangle \quad \text{for all } X \text{ tangent to } M, \end{aligned}$$

where Z is considered as an \mathbb{R}^3 -valued function on M , and $f_*(X)$ is identified with an element of \mathbb{R}^3 , or f is considered as an \mathbb{R}^3 -valued function on M . This equation is sometimes written simply

$$\langle dZ, df \rangle = 0 \quad \text{or} \quad dZ \cdot df = 0.$$

[Note: Sometimes X (or German \mathfrak{X}) is used to denote the *immersion* $X: M \rightarrow \mathbb{R}^3$, and this equation appears as $dZ \cdot dX = 0$.]

A vector field Z along an immersion $f: M \rightarrow \mathbb{R}^3$ will be called an **infinitesimal bending** of f if it satisfies equation (2'''). Clearly this equation will be satisfied by the variation vector field of a variation α which merely “preserves lengths up to first order”, equation (1) being, in fact, the analytic expression of this condition. Of course, we can always find infinitesimal bendings by taking the variation vector field Z of a bending α by means of Euclidean motions,

$$\alpha(t, p) = f(p) \cdot B(t) + v(t),$$

where $B(t) \in O(3)$ with $B(0) = I$, and $v(t) \in \mathbb{R}^3$ with $v(0) = 0$ [and $f(p) \cdot B(t)$ denotes the product of the 1×3 matrix $f(p)$ with the 3×3 matrix $B(t)$]. In this case we have

$$Z(p) = f(p) \cdot B'(0) + v'(0),$$

where $B'(t) \in \mathfrak{o}(3) = \{3 \times 3 \text{ skew-symmetric matrices}\}$. Conversely, if Z is an infinitesimal bending of f of the form

$$(3) \quad Z(p) = f(p) \cdot C + w, \quad C \in \mathfrak{o}(3),$$

then Z is the variation vector field of the bending α through Euclidean motions defined by

$$\alpha(t, p) = f(p) \cdot e^{tC} + tw.$$

So we will call an infinitesimal bending Z **trivial** if it is of the form (3). An immersion $f: M \rightarrow \mathbb{R}^3$ will be called **infinitesimally bendable** if there is a non-trivial infinitesimal bending of f ; otherwise it will be called **infinitesimally rigid**. (The word “rigid” is sometimes used to mean infinitesimally rigid, but unfortunately it is also sometimes sloppily used to mean unwarpable, or unbendable.)

Notice that the product of a vector by a skew-symmetric matrix,

$$(x, y, z) \cdot \begin{pmatrix} 0 & -a & -b \\ a & 0 & -c \\ b & c & 0 \end{pmatrix} = (ay + bz, -ax + cz, -bx - cy),$$

can also be written as a cross-product

$$(x, y, z) \times (c, -b, a) = (ay + bz, -ax + cz, -bx - cy).$$

So equation (3) can also be written

$$(3') \quad Z(p) = f(p) \times Y + w, \quad Y, w \in \mathbb{R}^3.$$

As an easy consequence of the triviality condition (3') we have

$$dZ(X) = df(X) \times Y \quad \text{all } X \text{ tangent to } M.$$

Now the same formula holds for an arbitrary infinitesimal bending Z , provided that we allow Y to vary:

3. LEMMA. If Z is an infinitesimal bending of $f: M \rightarrow \mathbb{R}^3$, then for each $p \in M$, there is a unique $Y(p) \in \mathbb{R}^3$ such that

$$dZ(X) = df(X) \times Y(p) \quad \text{for all } X \in M_p.$$

PROOF. Let $X_1, X_2 \in M_p$ be linearly independent. Since $\langle dZ(X_i), df(X_i) \rangle = 0$, there are certainly some vectors $Y_i \in \mathbb{R}^3$ with

$$dZ(X_i) = df(X_i) \times Y_i.$$

Moreover,

$$\begin{aligned} 0 &= \langle dZ(X_2), df(X_1) \rangle + \langle dZ(X_1), df(X_2) \rangle \\ &= \langle df(X_2) \times Y_2, df(X_1) \rangle + \langle df(X_1) \times Y_1, df(X_2) \rangle \\ &= \langle Y_2, df(X_1) \times df(X_2) \rangle - \langle Y_1, df(X_1) \times df(X_2) \rangle \\ &= \langle Y_2 - Y_1, df(X_1) \times df(X_2) \rangle. \end{aligned}$$

Thus $Y_2 - Y_1 \in df(M_p)$, so we can write

$$Y_2 - Y_1 = a df(X_1) + b df(X_2).$$

If we set

$$Y(p) = Y_2 - a df(X_1) = Y_1 + b df(X_2)$$

then we have

$$dZ(X_i) = df(X_i) \times Y_i = df(X_i) \times Y(p).$$

Uniqueness is obvious. ♦

The vector field $p \mapsto Y(p)$ of Lemma 3 is called the **(infinitesimal) rotation field** of the infinitesimal bending Z . We know that Y is constant when Z is trivial, and conversely,

4. LEMMA. If the rotation field Y of the infinitesimal bending Z is constant, then Z is trivial.

PROOF. By assumption, there is a vector $Y_0 \in \mathbb{R}^3$ with

$$dZ(X) = df(X) \times Y_0$$

for all X tangent to M . Let c be a curve in M , with $c(0) = p_0 \in M$. Then

$$\frac{dZ(c(t))}{dt} = dZ(c'(t)) = df(c'(t)) \times Y_0 = \frac{df(c(t))}{dt} \times Y_0.$$

Therefore

$$Z(c(t)) - Z(p_0) = [f(c(t)) - f(p_0)] \times Y_0,$$

or

$$Z(c(t)) = f(c(t)) \times Y_0 + w_0,$$

where $w_0 \in \mathbb{R}^3$ does not depend on c . So for all $p \in M$ we have

$$Z(p) = f(p) \times Y_0 + w_0,$$

and Z is trivial. ♦

At first sight it might seem that every bendable surface must also be infinitesimally bendable. As a matter of fact, we certainly do have

5. LEMMA. Let $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ be a bending, and let Z_t be the variation vector field of α at time t . If each Z_t is trivial, then the bending α is trivial.

PROOF. By definition,

$$(1) \quad Z_t(p) = \frac{d}{dt} \alpha(t, p),$$

and since each Z_t is trivial we have

$$(2) \quad Z_t(p) = \alpha(t, p) \times Y_t + w_t \quad Y_t, w_t \in \mathbb{R}^3.$$

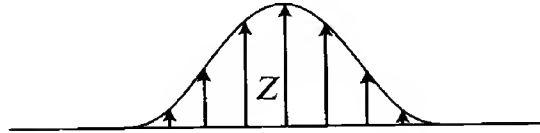
Then for all $p_1, p_2 \in M$ we have

$$\begin{aligned} \frac{d}{dt} |\alpha(t, p_1) - \alpha(t, p_2)|^2 &= 2 \langle \alpha(t, p_1) - \alpha(t, p_2), Z_t(p_1) - Z_t(p_2) \rangle \quad \text{by (I)} \\ &= 2 \langle \alpha(t, p_1) - \alpha(t, p_2), [\alpha(t, p_1) - \alpha(t, p_2)] \times Y_t \rangle \\ &= 0. \end{aligned}$$

So $|\alpha(t, p_1) - \alpha(t, p_2)|$ is constant in t . In particular, $|\alpha(t, p_1) - \alpha(t, p_2)| = |\alpha(0, p_1) - \alpha(0, p_2)|$. This implies that each α_t differs by a Euclidean motion from α_0 . ♦

Nevertheless, it is conceivable that $f: M \rightarrow \mathbb{R}^3$ is bendable, yet that every bending α of f has trivial variation vector field Z_0 at time $t = 0$, so that f is not infinitesimally bendable. No example of such a weird phenomenon is known, but there are also no positive results along this line, except for the obvious fact that if f is an analytic immersion which is analytically infinitesimally rigid, then f is analytically unbendable.

There are a couple of other surprising facts about infinitesimal bendings. First of all, there are non-trivial infinitesimal bendings Z of a plane which vanish outside a compact set. If our plane is the (x, y) -plane, we can choose Z to be $h \cdot \frac{\partial}{\partial z}$, where h is any C^∞ function vanishing outside a compact set. For any



tangent vector X of the (x, y) -plane we have

$$dZ(X) = (X(0), X(0), X(h)) = X(h) \cdot \frac{\partial}{\partial z},$$

so the infinitesimal rotation vector field of Z is

$$Y = X(h) \cdot \frac{\partial}{\partial z}.$$

Since Y is not constant, Z is non-trivial. Even more surprising, perhaps, is an immediate consequence of this fact: any surface containing a portion of a plane is infinitesimally bendable.

Notice that the infinitesimal bending Z constructed above is everywhere perpendicular to the surface $M = (x, y)$ -plane. This is essentially the only possibility:

6. LEMMA. (1) If Z is an infinitesimal bending of an open subset M of the (x, y) -plane, and Z is always tangential to the (x, y) -plane, then Z is trivial,

$$Z(p) = p \cdot C + v$$

for a 2×2 skew-symmetric matrix C , and $v \in (x, y)$ -plane.

(2) More generally, Z is an infinitesimal bending of $M \subset (x, y)$ -plane if and only if the tangential component $\mathbb{T}Z$ of Z is an infinitesimal bending, and hence trivial. In particular, any vector field Z normal to M is an infinitesimal bending.

(3) Let $M \subset \mathbb{R}^3$ be a surface and let Z be any infinitesimal bending of M which is everywhere normal to M . Then at every point $p \in M$ where $Z(p) \neq 0$, the second fundamental form $\Pi(p) = 0$. (So if $Z(p) \neq 0$ for all p in an open set $U \subset M$, then U lies in a plane.)

PROOF. (1) Let

$$Z(x, y) = (a(x, y), b(x, y)) = a(x, y) \frac{\partial}{\partial x} + b(x, y) \frac{\partial}{\partial y}.$$

Then

$$dZ \left(\frac{\partial}{\partial x} \right) = \frac{\partial a}{\partial x} \cdot \frac{\partial}{\partial x} + \frac{\partial b}{\partial x} \cdot \frac{\partial}{\partial y}, \quad dZ \left(\frac{\partial}{\partial y} \right) = \frac{\partial a}{\partial y} \cdot \frac{\partial}{\partial x} + \frac{\partial b}{\partial y} \cdot \frac{\partial}{\partial y}.$$

Thus

$$0 = \left\langle \frac{\partial}{\partial x}, dZ \left(\frac{\partial}{\partial x} \right) \right\rangle \implies \frac{\partial a}{\partial x} = 0$$

$$0 = \left\langle \frac{\partial}{\partial y}, dZ \left(\frac{\partial}{\partial y} \right) \right\rangle \implies \frac{\partial b}{\partial y} = 0$$

so we can write

$$a(x, y) = \bar{a}(y), \quad b(x, y) = \bar{b}(x).$$

Moreover,

$$0 = \left\langle \frac{\partial}{\partial x}, dZ \left(\frac{\partial}{\partial y} \right) \right\rangle + \left\langle \frac{\partial}{\partial y}, dZ \left(\frac{\partial}{\partial x} \right) \right\rangle \implies \bar{a}'(y) + \bar{b}'(x) = 0.$$

Since this is true for all x, y , the derivatives $\bar{a}'(y)$ and $\bar{b}'(x)$ must be constants. So we must have

$$a(x, y) = \bar{a}(y) = \alpha y + \beta$$

$$b(x, y) = \bar{b}(x) = -\alpha x + \delta.$$

Then

$$Z(x, y) = (x, y) \cdot \begin{pmatrix} 0 & -\alpha \\ \alpha & 0 \end{pmatrix} + (\beta, \delta).$$

This completes the proof of (1).

Now for any $M \subset \mathbb{R}^3$, with unit normal ν , consider a vector field

$$Z = \mathbb{T}Z + \phi \cdot \nu.$$

Then for $X \in M_p$ we have

$$\nabla'_X Z = \nabla_X \mathbb{T}Z + \text{II}(X, \mathbb{T}Z) \cdot \nu + X(\phi) \cdot \nu - \phi(p) d\nu(X).$$

So $0 = \langle \nabla'_X Z, X \rangle$ if and only if

$$(*) \quad \langle \nabla_X \mathbb{T}Z, X \rangle = \phi(p) \cdot \text{II}(X, X).$$

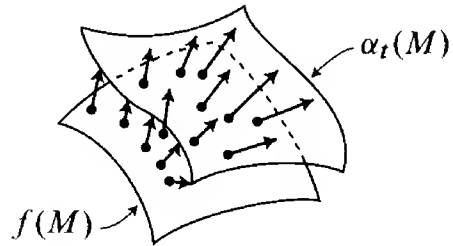
If $M \subset (x, y)$ -plane, then $\text{II}(X, X) = 0$ and $\text{II}(X, \mathbb{T}Z) = 0$, so $(*)$ says that Z is an infinitesimal bending if and only if $0 = \langle \nabla_X \mathbb{T}Z, X \rangle = \langle \nabla'_X \mathbb{T}Z, X \rangle$ for all X , so that $\mathbb{T}Z$ is an infinitesimal bending. This proves (2).

On the other hand, if Z is an infinitesimal bending with $\mathbb{T}Z = 0$, and $\phi(p) \neq 0$, then $(*)$ shows that $\text{II}(X, X) = 0$ for all $X \in M_p$. This proves (3). ♦

There also turns out to be a relationship between warpability and infinitesimal bendability, which at first sight seem to have nothing to do with each other.

7. LEMMA. Let Z be an infinitesimal bending of an immersion $f: M \rightarrow \mathbb{R}^3$. Define $\alpha_t: M \rightarrow \mathbb{R}^3$ by

$$\alpha_t(p) = f(p) + t \cdot Z(p),$$



where $Z(p)$ is considered as an element of \mathbb{R}^3 as usual. Then in a neighborhood of any point $p \in M$, the map α_t is an immersion for sufficiently small t , and the induced metric $\alpha_t^* \langle \cdot, \cdot \rangle$ on M is related to the metric $f^* \langle \cdot, \cdot \rangle$ by

$$[\alpha_t^* \langle \cdot, \cdot \rangle](X, Y) = [f^* \langle \cdot, \cdot \rangle](X, Y) + t^2 \langle dZ(X), dZ(Y) \rangle.$$

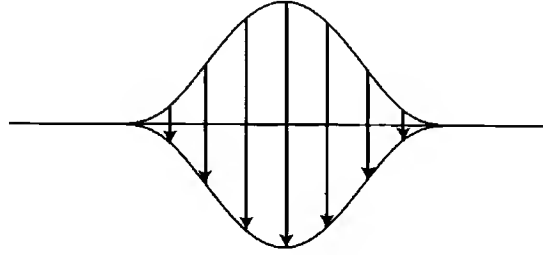
In particular, the metrics $\alpha_t^* \langle \cdot, \cdot \rangle$ and $\alpha_{-t}^* \langle \cdot, \cdot \rangle$ on M are the same.

PROOF. If X is a tangent vector on M , with $X = c'(0)$ for some curve c in M , then

$$\begin{aligned}\alpha_{t*}X &= \left. \frac{d}{ds} \right|_{s=0} \alpha_t(c(s)) \\ &= \left. \frac{d}{ds} \right|_{s=0} c(s) + t Z(c(s)) \\ &= c'(0) + t dZ(c'(0)) \\ &= X + t dZ(X).\end{aligned}$$

This immediately leads to the desired formula, and this formula shows that α_t is an immersion for small t , in any neighborhood of p on which Z is bounded. ♦

An illustration of this phenomenon is provided by the infinitesimal bending Z of the (x, y) -plane given previously. The map taking $p + tZ(p) \mapsto p - tZ(p)$



is the restriction of a Euclidean motion, namely reflection through the (x, y) -plane. But this is the *only* case in which this happens:

8. LEMMA. Let Z be a non-trivial infinitesimal bending of a surface $M \subset \mathbb{R}^3$ which is not part of a plane. Then for $t \neq 0$, the map

$$p + tZ(p) \mapsto p - tZ(p)$$

[which is an isometry by Lemma 7] is not the restriction of a Euclidean motion.

PROOF. If this map is the restriction of a length preserving map of \mathbb{R}^3 , then for all $p, q \in M$ we have

$$|p + tZ(p) - \{q + tZ(q)\}| = |p - tZ(p) - \{q - tZ(q)\}|,$$

that is,

$$|p - q + t\{Z(p) - Z(q)\}| = |p - q - t\{Z(p) - Z(q)\}|.$$

This implies that

$$(1) \quad \langle p - q, Z(p) - Z(q) \rangle = 0.$$

Without loss of generality, we may assume that M contains the point 0, and that $Z(0) = 0$. Then equation (1) gives

$$\langle p, Z(p) \rangle = 0 \quad p \in M,$$

from which we further deduce that

$$(2) \quad \langle p, Z(q) \rangle + \langle q, Z(p) \rangle = 0, \quad p, q \in M.$$

Since M is not contained in a plane, there are three linearly independent points $r_1, r_2, r_3 \in M$. Now if λ_i are numbers with $\sum_i \lambda_i r_i \in M$, then

$$\begin{aligned} \langle Z(\sum_i \lambda_i r_i), r_j \rangle &= -\langle \sum_i \lambda_i r_i, Z(r_j) \rangle && \text{by (2)} \\ &= -\sum_i \lambda_i \langle r_i, Z(r_j) \rangle \\ &= \sum_i \lambda_i \langle Z(r_i), r_j \rangle && \text{by (2)} \\ &= \langle \sum_i \lambda_i Z(r_i), r_j \rangle. \end{aligned}$$

This implies that

$$Z(\sum_i \lambda_i r_i) = \sum_i \lambda_i Z(r_i).$$

So Z is the restriction to M of a linear transformation T . Since a linear transformation is its own derivative, equation (2'') shows that

$$0 = \langle TX, Y \rangle + \langle X, TY \rangle$$

for all pairs of vectors X, Y which are in some M_p (when M_p is identified with a subspace of \mathbb{R}^3 in the usual way). Since M is not contained in a plane, there are three distinct subspaces $M_{p_1}, M_{p_2}, M_{p_3}$ (we regard these as vector subspaces of \mathbb{R}^3). Choose

$$X_3 \in M_{p_1} \cap M_{p_2}$$

$$X_2 \in M_{p_1} \cap M_{p_3}$$

$$X_1 \in M_{p_2} \cap M_{p_3}.$$

Then the X_i are linearly independent, and $0 = \langle TX_i, X_j \rangle + \langle X_i, TX_j \rangle$ for each i, j . This implies that $0 = \langle TX, Y \rangle + \langle X, TY \rangle$ for all $X, Y \in \mathbb{R}^3$. Thus T is skew-adjoint, and its matrix C is skew-symmetric. In other words,

$$Z(p) = p \cdot C \quad C \in \mathfrak{o}(3),$$

and hence Z is trivial. ♦

In order to obtain some deeper results about rigidity, we will find it useful to consider various \mathbb{R}^3 -valued differential forms on a surface M . Many of these forms will be defined in terms of other \mathbb{R}^3 -valued forms and functions on M by means of the inner product and cross product on \mathbb{R}^3 . If $f, g: M \rightarrow \mathbb{R}^3$ are two functions, then there is only one reasonable meaning for $f \times g$, namely the function

$$p \mapsto f(p) \times g(p) \in \mathbb{R}^3.$$

But if ω and η are \mathbb{R}^3 -valued 1-forms on M , then $\omega \times \eta$ might mean any of the following operations on tangent vectors:

$$\begin{aligned} X &\mapsto \omega(X) \times \eta(X) && \text{(a quadratic function)} \\ (X, Y) &\mapsto \omega(X) \times \eta(Y) && \text{(a bilinear function)} \\ (X, Y) &\mapsto \omega(X) \times \eta(Y) - \omega(Y) \times \eta(X) && \text{(a 2-form).} \end{aligned}$$

To distinguish these possibilities, we might write the last two as

$$\omega \overset{\otimes}{\times} \eta \quad \text{and} \quad \omega \overset{\wedge}{\times} \eta.$$

Since we shall, in fact, only be interested in the last case, we will introduce the simpler symbol \mathbf{x} and define

$$\omega \mathbf{x} \eta(X, Y) = \omega(X) \times \eta(Y) - \omega(Y) \times \eta(X).$$

The present situation is actually just a special case of the one already considered at the end of Chapter I.10; see pg. I.403 and especially Problems I.10-20 and 10-21. In general, if ω and η are \mathbb{R}^3 -valued forms of degree k and l , respectively, then we define a $(k+l)$ -form $\omega \mathbf{x} \eta$ by

$$\begin{aligned} &\omega \mathbf{x} \eta(X_1, \dots, X_k, X_{k+1}, \dots, X_{k+l}) \\ &= \frac{1}{k!l!} \sum_{\sigma \in S_{k+l}} \text{sgn } \sigma \cdot \omega(X_{\sigma(1)}, \dots, X_{\sigma(k)}) \times \eta(X_{\sigma(k+1)}, \dots, X_{\sigma(k+l)}). \end{aligned}$$

[Actually, since we will be dealing with a surface M , only the cases $k, l \leq 1$ are relevant.] In an exactly analogous way, we define $\omega \bullet \eta$, using the product

$v \cdot w = \langle v, w \rangle$ in \mathbb{R}^3 . Then we have (Problem I.10-21(a))

$$\begin{aligned} d(\omega \times \eta) &= d\omega \times \eta + (-1)^k \omega \times d\eta \\ d(\omega \cdot \eta) &= d\omega \cdot \eta + (-1)^k \omega \cdot d\eta. \end{aligned}$$

Notice that since \times is not commutative, $\omega \times \omega$ need not be zero. In fact, for a 1-form ω we have

$$\omega \times \omega(X, Y) = 2\omega(X) \times \omega(Y).$$

More generally,

$$\begin{aligned} \omega \times \eta &= (-1)^{kl+1} \eta \times \omega \\ \omega \cdot \eta &= (-1)^{kl} \eta \cdot \omega. \end{aligned}$$

It is also easy to see that the formula

$$v \cdot (w \times z) = -w \cdot (v \times z) \quad v, w, z \in \mathbb{R}^3$$

leads to the relation

$$\omega \cdot (\eta \times \lambda) = (-1)^{kl+1} \eta \cdot (\omega \times \lambda).$$

Pure notational fiddling would lead us to write $\omega \cdot (\eta \times \lambda)$ in the form **det**(ω, η, λ), which can be defined directly in an obvious way: if $\omega = (\omega^1, \omega^2, \omega^3)$ for ordinary 1-forms ω^i , and similarly for η and λ , then **det**(ω, η, λ) denotes

$$\det \begin{pmatrix} \omega^1 & \omega^2 & \omega^3 \\ \lambda^1 & \lambda^2 & \lambda^3 \\ \eta^1 & \eta^2 & \eta^3 \end{pmatrix},$$

where the determinant is expanded out as usual, with all multiplications being replaced by \wedge , and care being taken to write products in the correct order (namely, the same order as the columns they appear in). One easily checks, either from this definition, or from the alternative form $\omega \cdot (\eta \times \lambda)$, that

$$\begin{aligned} d \mathbf{det}(\omega, \eta, \lambda) &= \mathbf{det}(d\omega, \eta, \lambda) + (-1)^k \mathbf{det}(\omega, d\eta, \lambda) \\ &\quad + (-1)^{k+l} \mathbf{det}(\omega, \eta, d\lambda). \end{aligned}$$

We will frequently use the various formulas given here without specific comment.

Now consider an immersion $f: M \rightarrow \mathbb{R}^3$ of an oriented surface M , and the corresponding normal map $N: M \rightarrow \mathbb{R}^3$. If dA is the volume element of M

for the metric $f^*\langle \ , \ \rangle$, then we have the following identities among \mathbb{R}^3 -valued 2-forms on M :

$$\begin{aligned} \text{(I)} \quad & df \times df = 2N dA \\ \text{(II)} \quad & df \times dN = -2HN dA \\ \text{(III)} \quad & dN \times dN = 2KN dA. \end{aligned}$$

To prove these simple relations, pick vectors $X_1, X_2 \in M_p$ with (X_1, X_2) positively oriented. Now $df(X_i)$ is just $f_*(X_i)$, considered as an element of \mathbb{R}^3 . So

$$\begin{aligned} df \times df(X_1, X_2) &= 2 df(X_1) \times df(X_2) \\ &= 2N(p) \cdot \text{area of parallelogram spanned by } df(X_1), df(X_2) \\ &= 2N(p) \cdot dA(X_1, X_2). \end{aligned}$$

Moreover, if

$$\begin{aligned} dN(X_1) &= \alpha df(X_1) + \beta df(X_2) \\ dN(X_2) &= \gamma df(X_1) + \delta df(X_2), \end{aligned}$$

then

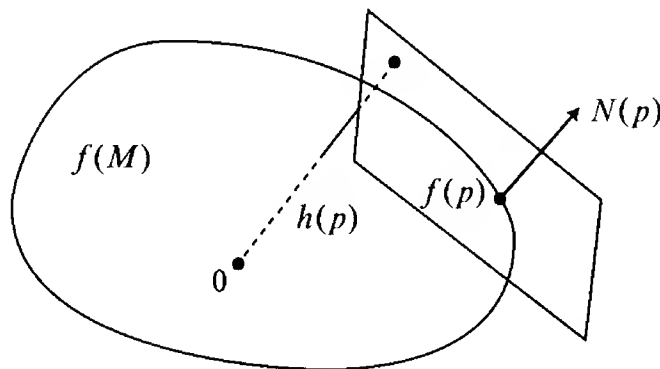
$$\begin{aligned} df \times dN(X_1, X_2) &= \delta df(X_1) \times df(X_2) - \alpha df(X_2) \times df(X_1) \\ &= (\alpha + \delta) df(X_1) \times df(X_2) \\ &= (\alpha + \delta)N(p) dA(X_1, X_2) \\ &= -2H(p)N(p) dA(X_1, X_2), \end{aligned}$$

and

$$\begin{aligned} dN \times dN(X_1, X_2) &= 2 dN(X_1) \times dN(X_2) \\ &= 2[\alpha df(X_1) + \beta df(X_2)] \times [\gamma df(X_1) + \delta df(X_2)] \\ &= 2(\alpha\delta - \beta\gamma) df(X_1) \times df(X_2) \\ &= 2K(p)N(p) dA(X_1, X_2). \end{aligned}$$

Naturally, all these formulas can be applied when $M \subset \mathbb{R}^3$, and f is just the inclusion map $i: M \rightarrow \mathbb{R}^3$. And, in fact, we shall usually apply them to imbedded, rather than immersed, surfaces. But it nevertheless seems conceptually easier always to regard M as an abstract surface sitting off in the void, so that f and N can be thought of simply as certain \mathbb{R}^3 -valued functions on M without worrying about the geometry they induce; most of the geometric information in question is already presented in formulas (I)–(III).

We will also need to recall the **support function** $h = -f \cdot N: M \rightarrow \mathbb{R}^3$, which is defined in Problem 3-7. As we saw, $h(p)$ is the signed distance from the origin to the tangent plane of $f(M)$ at $f(p)$; when f is an imbedding with



$f(M)$ star-shaped with respect to 0, and N is inward pointing, it is precisely this distance. This happens, in particular, when $f(M)$ is convex, and 0 lies inside it.

Now consider the \mathbb{R}^3 -valued 1-form α on M defined by

$$\alpha = (f \times N) \cdot df.$$

We have

$$\begin{aligned} d\alpha &= (df \times N) \cdot df + (f \times dN) \cdot df \\ &= -N \cdot (df \times df) + f \cdot (dN \times df) \quad \text{[the dots do not have to be bold since } f \text{ and } N \text{ are functions]} \\ &= -2dA - 2H(f \cdot N) dA \quad \text{by (I) and (II)} \\ &= -2dA + 2hH dA. \end{aligned}$$

If M is compact, then $\int_M d\alpha = \int_{\partial M} \alpha = 0$, so we obtain

$$(IV) \quad \text{area}(M) = \int_M hH dA.$$

Similarly, for the 1-form

$$\beta = (f \times N) \cdot dN,$$

we have

$$\begin{aligned} d\beta &= (df \times N) \cdot dN + (f \times dN) \cdot dN \\ &= -N \cdot (df \times dN) + f \cdot (dN \times dN) \\ &= 2H dA - 2hK dA \quad \text{by (II) and (III)}. \end{aligned}$$

Hence for compact M we have

$$(V) \quad \int_M H \, dA = \int_M hK \, dA.$$

Equations (IV) and (V) are sometimes called *Minkowski's formulas*.

As a first application of these formulas, we reprove a rigidity result which appeared a long time ago. The theorem that a compact surface with constant $K > 0$ must be a (standard) sphere in \mathbb{R}^3 can also be stated as follows: a sphere is unwarpable.* To prove this from our present formulas, we consider any compact surface $M \subset \mathbb{R}^3$ with constant $K > 0$. It is convex, by Hadamard's theorem, and we can assume that $0 \in \mathbb{R}^3$ lies in its interior, so that the support function h is always positive for the inward pointing N . There is no loss of generality in assuming that $K = k_1 k_2 = 1$. Since for $x > 0$ we always have

$$x + \frac{1}{x} \geq 2,$$

with equality only if $x = 1$, this implies that

$$H = \frac{1}{2} \left(k_1 + \frac{1}{k_1} \right) \geq 1,$$

with equality only if $k_1 = k_2$. So

$$\begin{aligned} \int_M h \, dA &= \int_M H \, dA && \text{by (V)} \\ &\geq \int_M dA \\ &= \int_M hH \, dA && \text{by (IV),} \end{aligned}$$

and consequently

$$\int_M h(1 - H) \, dA \geq 0.$$

Since $h > 0$ everywhere, and $1 - H \leq 0$ everywhere, it follows that $H = 1$ everywhere. This implies that $k_1 = k_2$ everywhere.

*Actually, the statement that a sphere is unwarpable is formally *stronger* than the statement that any compact surface isometric to a sphere is a sphere—see the footnote on page 171. But the complete equivalence of the two statements follows easily from the fact that any isometry of a sphere onto itself is the restriction of a Euclidean motion.

Similarly, we can reprove Theorem 5-3, and in a little greater generality.

9. PROPOSITION. The only star-shaped surfaces of constant mean curvature H are spheres.

PROOF. We still have $h > 0$, and there is no loss of generality in assuming that

$$1 = H = \frac{k_1 + k_2}{2}.$$

Then

$$K = k_1 k_2 = k_1(2 - k_1) = 2k_1 - (k_1)^2 \leq 1,$$

with equality only if $k_1 = 1$. Now

$$\begin{aligned} \int_M h dA &= \int_M dA && \text{by (IV)} \\ &= \int_M hK dA && \text{by (V),} \end{aligned}$$

so

$$\int_M h(1 - K) dA = 0.$$

Thus we must have $K = 1$ everywhere, hence $k_1 = 1$ everywhere, hence $k_2 = 1 = k_1$ everywhere. ♦

This proof is mainly a curiosity, since, as we showed in Chapter 9 (Addendum 2 or 3), a much stronger result actually holds. By considering another \mathbb{R}^3 -valued 1-form, however, we obtain a real theorem, one of the first in the subject.

10. THEOREM. Let $M \subset \mathbb{R}^3$ be any compact convex surface which does not contain a portion of a plane. Then M is infinitesimally rigid.

Remarks: (1) We have already pointed out that the conclusion fails if M does contain a portion of a plane.

(2) We could replace convexity with the hypothesis $K \geq 0$ (see the remark after Theorem 2-11 or Proposition 7-32).

PROOF. For simplicity, we first consider the case where $K > 0$ everywhere. Let Z be an infinitesimal bending of the inclusion map $f: M \rightarrow \mathbb{R}^3$, and let Y be its rotation field, so that

$$dZ(X_p) = df(X_p) \times Y(p)$$

for all $X_p \in M_p$. This relation can be written in terms of the \mathbb{R}^3 -valued 1-forms dZ and df , and \mathbb{R}^3 -valued function Y , as

$$dZ = df \times Y.$$

Now

$$(1) \quad 0 = d(dZ) = -df \times dY.$$

This means that for $X_1, X_2 \in M_p$ we have

$$(2) \quad df(X_1) \times dY(X_2) - df(X_2) \times dY(X_1) = 0.$$

Taking the dot product of this equation with $df(X_1)$ and $df(X_2)$, we find that $dY(X_i)$ lies in the plane spanned by $df(X_1)$ and $df(X_2)$, which is nothing but the tangent plane M_p moved over to the origin. In other words, we can consider dY as a map $dY: M_p \rightarrow M_p$.

Now choose a moving frame X_1, X_2 in a neighborhood of p . We can write

$$(3) \quad \begin{aligned} dY(X_1) &= \alpha df(X_1) + \beta df(X_2) \\ dY(X_2) &= \gamma df(X_1) + \delta df(X_2) \end{aligned}$$

for some functions $\alpha, \beta, \gamma, \delta$. Equation (1) implies that

$$(4) \quad \alpha + \delta = 0.$$

Remembering that f is just the inclusion map, so that X_1, X_2 are vector fields on M , we can write

$$(5) \quad \begin{aligned} 0 &= d(dY)(X_1, X_2) = X_1(dY(X_2)) - X_2(dY(X_1)) - dy([X_1, X_2]) \\ &= \gamma \nabla'_{X_1} X_1 + \delta \nabla'_{X_1} X_2 - \alpha \nabla'_{X_2} X_1 - \beta \nabla'_{X_2} X_2 \\ &\quad + \text{something tangent to } M. \end{aligned}$$

Taking the inner product with N , and using (4), we get

$$\begin{aligned} 0 &= \gamma \Pi(X_1, X_1) - 2\alpha \Pi(X_1, X_2) - \beta \Pi(X_2, X_2) \\ &= \gamma l - 2\alpha m - \beta n, \quad \text{say.} \end{aligned}$$

In particular, suppose that we choose X_1, X_2 to be principal vectors at some point p , so that at p we have $m = 0$. Then our equation is simply

$$(6) \quad 0 = \gamma l - \beta n.$$

Since $K = ln > 0$, this shows that γ and β have the same sign. So

$$0 \leq \beta\gamma \quad \text{and} \quad 0 = \beta\gamma \quad \text{only if} \quad \beta = \gamma = 0.$$

Hence at each point p we have

$$(7) \quad 0 \leq \alpha^2 + \beta\gamma = -\det dY$$

with equality only if $\alpha = \beta = \gamma = 0 \implies dY = 0$.

Consider the 1-form

$$\omega = (f \times Y) \cdot dY$$

(closely related to the 1-form β considered previously). We have

$$\begin{aligned} d\omega &= (df \times Y) \cdot dY + (f \times dY) \cdot dY \\ &= -Y \cdot (df \times dY) + f \cdot (dY \times dY) \\ &= 0 + f \cdot (dY \times dY) \quad \text{by (I).} \end{aligned}$$

But we also have

$$dY \times dY = 2(\det dY)N dA,$$

by the very same argument which proved formula (III). Hence

$$d\omega = 2h(\det dY) dA.$$

So for our compact manifold M we have the *integral formula of Blaschke*:

$$(*) \quad \int_M h(\det dY) dA = 0.$$

Since $h > 0$, and $\det dY \leq 0$ by (7), we must have $\det dY = 0$ everywhere. Then (7) also shows that $dY = 0$ everywhere. Therefore Y is constant. So Z is trivial by Lemma 4.

Now we consider the case where $K \geq 0$, but M contains no portion of a plane, so that the planar points of M are nowhere dense. At a parabolic point we have $n = 0$ and $l \neq 0$, say. Equation (6) then shows that $\gamma = 0$. Hence we still have

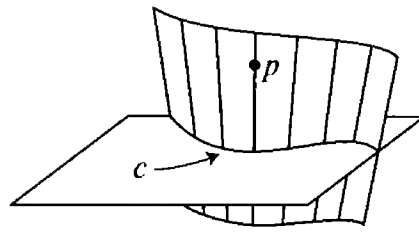
$$0 \leq \alpha^2 + \beta\gamma = -\det dY.$$

Thus we have $\det dY \leq 0$ at all non-planar points, which implies that $\det dY \leq 0$ everywhere. So we can still conclude from (*) that $\det dY = 0$ everywhere. We want to show that consequently $dY = 0$ everywhere; it obviously suffices to show that $dY(p) = 0$ when p is a parabolic point.

If the parabolic point p lies in the closure of $\{q \in M : K(q) > 0\}$, then clearly $dY(p) = 0$. So consider a parabolic point p which has a neighborhood on which $K = 0$. By Proposition 5-4 and Corollary 5-6, M contains a ruled surface

$$(s, t) \mapsto c(s) + td(s), \quad |d| = 1 \implies \langle d, d' \rangle = 0$$

around p such that the ruling through p has its endpoints in the closure of $\{q \in M : K(q) > 0\}$, and consequently $dY = 0$ at the endpoints of this ruling. We can choose the curve c to be the intersection of the ruled surface with a



plane perpendicular to the ruling through p . So if X_1, X_2 are the coordinate vector fields

$$\begin{aligned} X_1(s, t) &= c'(s) + td'(s) \\ X_2(s, t) &= d(s), \end{aligned}$$

then along the ruling through p we have

$$\langle X_1, X_2 \rangle = \langle c' + td', d \rangle = \langle c', d \rangle = 0.$$

Now X_2 is the principal vector with principal curvature $n = 0$. So along the ruling through p , the vector field X_1 is the other principal vector, with principal curvature $l \neq 0$. So we have [by (6)]

$$(8) \quad \gamma = 0 \quad \text{along the ruling through } p.$$

Since we have $0 = -\det Y = \alpha^2 + \beta\gamma$ everywhere, we also have

$$(9) \quad \alpha = 0 \quad \text{along the ruling through } p.$$

So equations (3) and (4) give

$$\begin{aligned} dY(X_1) &= \beta df(X_2) \\ dY(X_2) &= 0 \end{aligned} \quad \text{along the ruling through } p.$$

Then equation (5) becomes simply

$$\begin{aligned} 0 &= 0 - X_2(\beta df(X_2)) - 0 \\ &= -X_2(\beta) df(X_2) - \beta \nabla'_{X_2} X_2 \\ &= -X_2(\beta) df(X_2). \end{aligned}$$

Thus $0 = X_2(\beta)$ along the ruling through p , so that β is constant on this ruling. But $\beta = 0$ at the endpoints, since $dY = 0$ at the endpoints. It follows that also $\beta = 0$ at p ; together with (8) and (9) we now have $dY(p) = 0$. ♦

At this point it might be nice to have some non-trivial examples of surfaces which *are* infinitesimally bendable. Consider a surface given as a graph,

$$f(x, y) = (x, y, u(x, y));$$

we introduce the standard notation

$$\begin{aligned} p &= u_1 = \frac{\partial u}{\partial x}, & q &= \frac{\partial u}{\partial y} \\ r &= \frac{\partial^2 u}{\partial x^2}, & s &= \frac{\partial^2 u}{\partial x \partial y}, & t &= \frac{\partial^2 u}{\partial y^2}. \end{aligned}$$

Suppose that $Z = (_, _, \zeta)$ is an infinitesimal bending, with rotation field $Y = (\alpha, \beta, \psi)$. Then

$$\begin{aligned} (_, _, \zeta_1) &= dZ \left(\frac{\partial}{\partial x} \right) = df \left(\frac{\partial}{\partial x} \right) \times (\alpha, \beta, \psi) \\ &= (1, 0, p) \times (\alpha, \beta, \psi) \\ &= (_, _, \beta) \end{aligned}$$

and similarly

$$(_, _, \zeta_2) = (_, _, -\alpha).$$

So Y must be of the form

$$Y = (-\zeta_2, \zeta_1, \psi).$$

Using equation (2) in the proof of Theorem 10, we see that we must have

$$(1, 0, p) \times (-\zeta_{22}, \zeta_{12}, \psi_2) - (0, 1, q) \times (-\zeta_{12}, \zeta_{11}, \psi_1) = 0,$$

which is equivalent to

$$(*) \quad \begin{cases} \psi_1 = q\zeta_{11} - p\zeta_{12} \\ \psi_2 = -q\zeta_{12} + p\zeta_{22}. \end{cases}$$

Conversely, if ψ_i, ζ_{ij} satisfy these equations, and we define the vector-valued 1-form W by

$$W(X) = df(X) \times (-\zeta_2, \zeta_1, \psi),$$

then W will satisfy $dW = 0$, so on any simply-connected portion of the (x, y) -plane there will be Z with $dZ = W$. Now equations $(*)$ can be solved for ψ if and only if

$$(q\zeta_{11} - p\zeta_{12})_2 = (-q\zeta_{12} + p\zeta_{22})_1,$$

which leads to an equation for ζ :

$$(**) \quad r\zeta_{22} - 2s\zeta_{12} + t\zeta_{11} = 0.$$

As a particular case, we consider the paraboloid $u(x, y) = \frac{1}{2}(x^2 + y^2)$. We obtain the equation

$$\zeta_{11} + \zeta_{22} = 0,$$

whose solutions are the real part of any entire function on $\mathbb{C} = \mathbb{R}^2$. Thus there are non-trivial infinitesimal bendings of $\{(x, y, u(x, y))\}$, which is a complete convex surface.

As we have already pointed out, it is not known whether infinitesimal rigidity generally implies unbendability. But we *can* deduce this further property in the special situation considered in Theorem 10.

11. COROLLARY. Let $M \subset \mathbb{R}^3$ be a compact convex surface with $K > 0$ everywhere. Then M is unbendable.

PROOF. Let $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ be any bending of the inclusion map $i: M \rightarrow \mathbb{R}^3$. Then all $\bar{\alpha}(t)(M)$ have $K > 0$ everywhere, so all $\bar{\alpha}(t)(M)$ are infinitesimally rigid, by Theorem 10. So the variation vector field Z_t of α at time t is trivial. Then by Lemma 5, the bending α is trivial. ♦

In this corollary, the case $K \geq 0$ eluded us, but we aren't going to worry very much about it, because we are now going to prove a much better result anyway, the famous theorem of Cohn-Vossen that any convex surface is unwarpable. This result is the uniqueness part of Weyl's Problem, mentioned in the previous chapter; the present proof stems from the work of Herglotz.

12. THEOREM (COHN-VOSSEN). If $M \subset \mathbb{R}^3$ is a compact convex surface, then M is unwarpable.

PROOF. As in the proof of Theorem 11, for simplicity we first consider the case where $K > 0$ everywhere. So we consider two imbeddings $f, \bar{f}: M \rightarrow \mathbb{R}^3$ with $f^*\langle \cdot, \cdot \rangle = \bar{f}^*\langle \cdot, \cdot \rangle$, such that the curvature $K = \bar{K}$ for this metric is > 0 everywhere. Let N, \bar{N} be the inward pointing normals, for the convex surfaces $f(M)$ and $\bar{f}(M)$, and orient these surfaces so that N and \bar{N} are the normals determined by the orientations. We can assume that M has an orientation which makes both maps $f: M \rightarrow f(M)$ and $\bar{f}: M \rightarrow \bar{f}(M)$ orientation preserving (by composing \bar{f} with a reflection if necessary). For each $p \in M$ we have two subspaces $df(M_p), d\bar{f}(M_p) \subset \mathbb{R}^3$, which are just $f_*(M_p)$ and $\bar{f}_*(M_p)$ moved over to the origin. So we can consider

$$\iota = d(f \circ \bar{f}^{-1}): d\bar{f}(M_p) \rightarrow df(M_p).$$

The magic 1-form which we want to consider is

$$\omega = (f \times N) \cdot (\iota \circ d\bar{N}).$$

Figuring out $d\omega$ will be quite a bit harder than in the previous theorem. First we will get an expression for ω in terms of moving frames. We choose a moving frame X_1, X_2 on M which is orthonormal for $f^*\langle \cdot, \cdot \rangle = \bar{f}^*\langle \cdot, \cdot \rangle$, and let θ^1, θ^2 and $\omega_1^2 = -\omega_2^1$ be its dual forms and connection forms. We can consider (f_*X_1, f_*X_2, N) and $(\bar{f}_*X_1, \bar{f}_*X_2, \bar{N})$ as adapted orthonormal moving frames on $f(M)$ and $\bar{f}(M)$; let ψ_1^3, ψ_2^3 be f^* of the corresponding forms on $f(M)$, and define $\bar{\psi}_1^3, \bar{\psi}_2^3$ similarly. Then for X tangent to M we have

$$\begin{aligned} d\bar{N}(X) &= \bar{\psi}_1^3(X) \cdot d\bar{f}(X_1) + \bar{\psi}_2^3(X) \cdot d\bar{f}(X_2) \\ (1) \quad &\Downarrow \\ \iota \circ d\bar{N}(X) &= \bar{\psi}_1^3(X) \cdot df(X_1) + \bar{\psi}_2^3(X) \cdot df(X_2). \end{aligned}$$

We will also express $f(p)$ as a linear combination

$$(2) \quad f(p) = y_1(p) \cdot df(X_1(p)) + y_2(p) \cdot df(X_2(p)) + y_3(p) \cdot N(p),$$

where, in particular,

$$(3) \quad y_3 = f \cdot N = -h.$$

Then for $X \in M_p$ we have

$$\begin{aligned}
 \omega(X) &= [f(p) \times N(p)] \cdot [\bar{\psi}_3^1(X) \cdot df(X_1) + \bar{\psi}_3^2(X) \cdot df(X_2)] && \text{by (1)} \\
 &= [y_1(p) \cdot df(X_1) \times N(p) + y_2(p) \cdot df(X_2) \times N(p)] \\
 &\quad \cdot [\bar{\psi}_3^1(X) \cdot df(X_1) + \bar{\psi}_3^2(X) \cdot df(X_2)] && \text{by (2)} \\
 &= y_1(p) \cdot \bar{\psi}_3^2(X) - y_2(p) \cdot \bar{\psi}_3^1(X),
 \end{aligned}$$

and consequently

$$(4) \quad \omega = -y_1 \bar{\psi}_2^3 + y_2 \bar{\psi}_1^3.$$

Now equation (2) implies that for $X \in M_p$ we have

$$\begin{aligned}
 df(X) &= dy_1(X) \cdot df(X_1(p)) + y_1(p) \nabla'_{f_* X} f_* X_1 + \dots \\
 &= dy_1(X) \cdot df(X_1(p)) + y_1(p) \omega_1^2(X) \cdot df(X_2(p)) \\
 &\quad + y_1(p) \psi_1^3(X) N(p) + \dots \\
 &= [dy_1(X) + y_2(p) \omega_2^1(X) + y_3(p) \psi_3^1(X)] df(X_1) + \dots.
 \end{aligned}$$

But also

$$df(X) = \theta^1(X) \cdot df(X_1) + \theta^2(X) \cdot df(X_2).$$

Hence we have

$$\begin{aligned}
 (5) \quad dy_1 &= \theta^1 + y_2 \omega_1^2 + y_3 \psi_1^3 \\
 dy_2 &= \theta^2 + y_1 \omega_2^1 + y_3 \psi_2^3.
 \end{aligned}$$

Now we can compute

$$\begin{aligned}
 d\omega &= d(-y_1 \bar{\psi}_2^3 + y_2 \bar{\psi}_1^3) && \text{by (4)} \\
 &= -(\theta^1 + y_2 \omega_1^2 + y_3 \psi_1^3) \wedge \bar{\psi}_2^3 + y_1 (\bar{\psi}_1^3 \wedge \omega_2^1) \\
 &\quad + (\theta^2 + y_1 \omega_2^1 + y_3 \psi_2^3) \wedge \bar{\psi}_1^3 - y_2 (\bar{\psi}_2^3 \wedge \omega_1^2) \\
 &\quad \text{by (5) and the structural equations} \\
 &= -\theta^1 \wedge \bar{\psi}_2^3 + \theta^2 \wedge \bar{\psi}_1^3 - y_3 (\psi_1^3 \wedge \bar{\psi}_2^3 - \psi_2^3 \wedge \bar{\psi}_1^3) \\
 &= -\{(\bar{l}_{11} + \bar{l}_{22}) + y_3 (l_{11} \bar{l}_{22} + \bar{l}_{11} l_{22} - 2l_{12} \bar{l}_{12})\} dA,
 \end{aligned}$$

where

$$l_{ij} = \Pi_f(X_i, X_j), \quad \bar{l}_{ij} = \Pi_{\bar{f}}(X_i, X_j), \quad dA = \theta^1 \wedge \theta^2.$$

Now one has to observe that

$$\begin{aligned} l_{11}\bar{l}_{22} - 2l_{12}\bar{l}_{12} + l_{11}\bar{l}_{12} &= (l_{11}l_{22} - l_{12}^2) + (\bar{l}_{11}\bar{l}_{22} - \bar{l}_{12}^2) \\ &\quad - \det \begin{pmatrix} \bar{l}_{11} - l_{11} & \bar{l}_{12} - l_{12} \\ \bar{l}_{12} - l_{12} & \bar{l}_{22} - l_{22} \end{pmatrix} \\ &= 2K - \det(d\bar{N} - dN), \end{aligned}$$

where we now regard dN and $d\bar{N}$ as maps $dN, d\bar{N}: M_p \rightarrow M_p$. We obtain finally

$$d\omega = -\{2\bar{H} - h(2K - \det(d\bar{N} - dN))\} dA.$$

So for our compact M we have

$$(6) \quad 2 \int_M \bar{H} dA - 2 \int_M hK dA = - \int_M h \det(d\bar{N} - dN) dA.$$

Using formula (V), we obtain the *Herglotz integral formula*:

$$(7) \quad 2 \int_M \bar{H} - H dA = - \int_M h \det(d\bar{N} - dN) dA.$$

[Note that formula (V) also follows from (6) by taking $f = \bar{f}$.] Now we need an algebraic

13. LEMMA. Let A and B be two self-adjoint linear transformations on \mathbb{R}^2 which are positive semi-definite (i.e., have eigenvalues ≥ 0). Suppose that $\det A = \det B$. Then

$$\det(A - B) \leq 0.$$

Moreover, if A and B are positive definite, then equality holds only if $A = B$; and if A and B are positive semi-definite, then equality holds only if A and B are proportional.

PROOF. Consider A and B as symmetric matrices, and suppose first that A is positive definite. Since A is self-adjoint, there is an orthogonal matrix P , with transpose P^t , satisfying

$$PAP^t = PAP^{-1} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad \lambda_1, \lambda_2 > 0.$$

If we set

$$Q = CP, \quad C = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} \end{pmatrix},$$

then

$$(a) \quad Q A Q^t = C P A P^t C = I.$$

Now consider $Q B Q^t$. It is also symmetric, so there is an orthogonal R with

$$(b) \quad (R Q) B (R Q)^t = R (Q B Q^t) R^t = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}, \quad \mu_1, \mu_2 > 0.$$

Moreover, equation (a) gives

$$(c) \quad (R Q) A (R Q)^t = R (Q A Q^t) R^{-1} = I.$$

[We have simply reproved the well-known result that two positive definite quadratic forms can be simultaneously diagonalized.] So for $S = R Q$ we have

$$\begin{aligned} (\det S)^2 \det B &= \mu_1 \mu_2 && \text{by (b),} \\ (\det S)^2 \det A &= 1 && \text{by (c).} \end{aligned}$$

If $\det A = \det B$, then

$$\mu_1 \mu_2 = 1.$$

Moreover,

$$(\det S)^2 \det(A - B) = \det \begin{pmatrix} 1 - \mu_1 & 0 \\ 0 & 1 - \mu_2 \end{pmatrix} = 2 - (\mu_1 + \mu_2).$$

Since for $x > 0$ we always have

$$x + \frac{1}{x} \geq 2, \quad \text{with equality only if } x = 1,$$

it follows that

$$2 - (\mu_1 + \mu_2) \leq 0,$$

with equality only if $\mu_1 = \mu_2 = 1 \implies A = B$.

Now suppose that A and B are positive semi-definite, with $A \neq 0$, say. We can now obtain Q with

$$(a') \quad Q A Q^t = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

and R with

$$(b') \quad (R Q) B (R Q)^t = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} \quad \mu_1, \mu_2 \geq 0, \quad \mu_1 \mu_2 = 0.$$

As before, we also have

$$(c') \quad (RQ)A(RQ)^t = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

So for $S = RQ$ we have

$$(\det S)^2 \det(A - B) = \det \begin{pmatrix} 1 - \mu_1 & 0 \\ 0 & -\mu_2 \end{pmatrix} = -\mu_2 + \mu_1\mu_2 = -\mu_2.$$

Thus $\det(A - B) \leq 0$, with equality only if $\mu_2 = 0 \implies B = \mu_1 \cdot A$. **Q.E.D.**

Applying the Lemma to the positive definite maps $dN, d\bar{N}: M_p \rightarrow M_p$, with the same determinant $K(p)$, we conclude from equation (7) that

$$\int_M \bar{H} dA - \int_M H dA \geq 0.$$

But we can interchange f and \bar{f} in this inequality to obtain

$$\int_M H dA - \int_M \bar{H} dA \geq 0.$$

Hence

$$\int_M H dA = \int_M \bar{H} dA.$$

Then equation (7) gives

$$(*) \quad \int_M h \det(d\bar{N} - dN) dA = 0.$$

Now $(*)$ implies that $\det(d\bar{N} - dN) = 0$ everywhere. Then Lemma 13 implies that $d\bar{N} = dN$ everywhere. So the fundamental theorem of surface theory implies that f and \bar{f} differ by a Euclidean motion.

Now we consider the case $K \geq 0$. We can still obtain $(*)$ and thus conclude that $\det(d\bar{N} - dN) = 0$ everywhere. We have to show that $d\bar{N}(p) = dN(p)$ for points p with $K(p) = 0$, and it is only necessary to consider points p with $K = 0$ in a whole neighborhood of p . If $f(p)$ and $\bar{f}(p)$ are both planar points, there is nothing to prove. So suppose that $f(p)$, say, is a parabolic point. Then, as in the previous proof, the point $f(p)$ is on some line segment $\Gamma \subset f(M)$, whose endpoints Q_1, Q_2 are in the closure of the set where $K > 0$. Let $\bar{\Gamma} \subset \bar{f}(M)$

be the image of Γ under the isometry $\bar{f} \circ f^{-1}: f(M) \rightarrow \bar{f}(M)$. Then $\bar{\Gamma}$ is a geodesic in $\bar{f}(M)$. Now Γ is also an asymptotic curve, $\Pi(X, X) = 0$ for tangent vectors X pointing along Γ . The last part of Lemma 13 then shows that we must have $\bar{\Pi}(Y, Y) = 0$ for tangent vectors Y pointing along $\bar{\Gamma}$. So $\bar{\Gamma}$ has normal curvature $\bar{\kappa}_n = 0$; since $\bar{\Gamma}$ has geodesic curvature $\bar{\kappa}_g = 0$, it follows that $\bar{\Gamma}$ has curvature $\bar{\kappa} = \sqrt{\bar{\kappa}_n^2 + \bar{\kappa}_g^2} = 0$. Hence $\bar{\Gamma}$ is also a straight line segment, with endpoints \bar{Q}_1, \bar{Q}_2 , say. Lemma 5-5 says that the non-zero principal curvature k along Γ is of the form

$$k(s) = \frac{1}{As + B},$$

where $k(s)$ is the value of k at the point on Γ at distance s from $f(p)$. In particular, k cannot approach zero at Q_1 or Q_2 , so Q_1 and Q_2 are not planar points. Since Q_1, Q_2 are in the closure of the set where $K > 0$ we have $dN(Q_i) = d\bar{N}(\bar{Q}_i)$, so \bar{Q}_1, \bar{Q}_2 are also not planar points. So by Corollary 5-6, $\bar{f}(p)$ is not a planar point. Thus the non-zero principal curvature \bar{k} along $\bar{\Gamma}$ is of the form

$$\bar{k}(s) = \frac{1}{\bar{A}s + \bar{B}},$$

where s now measures the distance from $\bar{f}(p)$. But since $dN(Q_i) = d\bar{N}(\bar{Q}_i)$, we have $k(Q_i) = \bar{k}(\bar{Q}_i)$. It follows that $A = \bar{A}$ and $B = \bar{B}$. Hence $d\bar{N}(p) = dN(p)$. ♦

For later use, we insert here a form of the Herglotz integral formula for compact surfaces-with-boundary.

14. LEMMA. Let M be a compact oriented surface with boundary ∂M , and let $f, \bar{f}: M \rightarrow \mathbb{R}^3$ be immersions with $f^*\langle \cdot, \cdot \rangle = \bar{f}^*\langle \cdot, \cdot \rangle$. Let $N, \bar{N}: M \rightarrow \mathbb{R}^3$ be the normals determined by the orientation, let dA be the volume form on $(M, f^*\langle \cdot, \cdot \rangle)$, and let ds be the volume form on ∂M . For $p \in \partial M$ let $\mathbf{t}(p), \mathbf{u}(p)$ be the first two vectors of the Darboux frame at $f(p)$ for the curve $f(\partial M)$ on $f(M)$; we regard \mathbf{t} and \mathbf{u} as elements of \mathbb{R}^3 , as usual. Let κ_n and τ_g be the normal curvature and geodesic torsion for this curve, and let $\bar{\kappa}_n$ and $\bar{\tau}_g$ be the corresponding quantities for the curve $\bar{f}(\partial M)$ on $\bar{f}(M)$. Then

$$\begin{aligned} \int_{\partial M} (\bar{\tau}_g - \tau_g) \langle f, \mathbf{t} \rangle + (\kappa_n - \bar{\kappa}_n) \langle f, \mathbf{u} \rangle ds \\ = \int_M h \det(d\bar{N} - dN) dA + 2 \int_M \bar{H} - H dA. \end{aligned}$$

PROOF. We consider the 1-form

$$\omega = (f \times N) \cdot (\iota \circ d\bar{N})$$

of the previous proof, for which we have

$$(1) \quad d\omega = -\{2\bar{H} - h(2K - \det(d\bar{N} - dN))\} dA.$$

If \mathbf{s} is the unit tangent vector of the curve ∂M on $(M, f^*\langle \cdot, \cdot \rangle)$, so that $df(\mathbf{s}) = \mathbf{t}$, then

$$dN(\mathbf{s}) = -\kappa_n \mathbf{t} - \tau_g \mathbf{u},$$

by definition of κ_n and τ_g . Similarly,

$$(\iota \circ d\bar{N})(\mathbf{s}) = -\bar{\kappa}_n \mathbf{t} - \bar{\tau}_g \mathbf{u}.$$

Therefore

$$(2) \quad \begin{aligned} \omega(\mathbf{s}) &= f \cdot N \times (\iota \circ d\bar{N})(\mathbf{s}) = f \cdot N \times [-\bar{\kappa}_n \mathbf{t} - \bar{\tau}_g \mathbf{u}] \\ &= f \cdot (\bar{\tau}_g \mathbf{t} - \bar{\kappa}_n \mathbf{u}) \\ &= \bar{\tau}_g \langle f, \mathbf{t} \rangle - \bar{\kappa}_n \langle f, \mathbf{u} \rangle. \end{aligned}$$

Substituting (1) and (2) into Stokes' Theorem,

$$\int_{\partial M} \omega = \int_M d\omega,$$

we obtain

$$(3) \quad \begin{aligned} \int_{\partial M} \bar{\tau}_g \langle f, \mathbf{t} \rangle - \bar{\kappa}_n \langle f, \mathbf{u} \rangle ds &= -2 \int_M \bar{H} dA + 2 \int_M hK dA \\ &\quad + \int_M h \det(d\bar{N} - dN) dA. \end{aligned}$$

We cannot use formula (V) for $\int_M hk dA$, since M is not compact. But choosing $\bar{f} = f$ in (3) we obtain

$$\int_{\partial M} \tau_g \langle f, \mathbf{t} \rangle - \kappa_n \langle f, \mathbf{u} \rangle ds = -2 \int_M H dA + 2 \int_M hK dA.$$

Substituting this into (3), we obtain the desired result. \blacklozenge

The proofs of Theorems 10 and 12 have a formal correspondence which is even more complete than their superficial resemblance. Indeed, suppose that the two imbeddings f, \bar{f} of Theorem 12 are part of a variation α of f , with variation vector field Z . Then each $\bar{\alpha}(t)$ has a normal field N_t , and the integrand $h \det(d\bar{N}_t - dN)$ of the Herglotz integral formula can be expanded in powers of t ; the terms up to second order in t turn out to be exactly the integrand in Blaschke's integral formula.

More to the point, perhaps, is the fact that the proofs of Theorems 10 and 12 are equally mysterious. They depend on discovering 1-forms ω for which $d\omega = (\text{something interesting}) \cdot dA$; these 1-forms ω are suggested by the geometry in only the vaguest way, and one simply has to carry out the computations explicitly to see what $d\omega$ really is. In this connection, however, the following may be mentioned. The requirement that \bar{f} be an imbedding with the same metric as f is a system of partial differential equations (in 2 variables) for the components of \bar{f} . (The requirement that α be a bending of f is an even more complicated equation in 3 variables, and the basic aim of introducing infinitesimal bendings is to reduce the problem to one in only 2 variables. This "linearization" of the problem leads to a system of linear partial differential equations.) Theorems 10 and 12 may be regarded as uniqueness theorems for partial differential equations on M . As Stoker {1} points out, "the proofs of uniqueness theorems for boundary-value problems involving other partial differential equations also usually require the invention of special tricks and devices, above all if the problems are nonlinear, and such devices commonly involve integrals over the domains in question (e.g., energy integrals in problems having their origin in mathematical physics)."

This is perhaps an opportune moment to describe briefly the original, quite geometric, proofs of these rigidity results. Theorem 10 was first proved by Liebmann, and the crux of his proof was the following observation. Let Z be an infinitesimal bending of $M \subset \mathbb{R}^3$. Regarding each $Z(p)$ as an element of \mathbb{R}^3 , we obtain a surface $Z = \{Z(p) : p \in M\} \subset \mathbb{R}^3$. Of course, this may not really be an immersed surface; indeed we hope to prove that it contains only one point when M is compact with $K > 0$. Liebmann showed that at points $p \in M$ where $p \mapsto Z(p)$ is an immersion, the curvature of Z is < 0 at $Z(p)$ when the curvature of M is > 0 at p . This immediately shows that $p \mapsto Z(p)$ cannot be an immersion everywhere when M is compact with $K > 0$, since Z would then be a compact surface with $K < 0$. Liebmann showed that even when Z has singularities, it is nevertheless true that if Z is not a point, then Z has the character of a surface of negative curvature, in the sense that no point q of Z has a support plane (a plane containing q and all points of Z on one side of it);

this property again contradicts the compactness of Z . Since Liebmann's proof involves the investigation of singularities, it is hardly surprising that it works only in the analytic case.

Cohn-Vossen's proof was also originally restricted to the analytic case, and is quite similar to the proof of Hopf's Theorem (Theorem 9-33) on surfaces of constant mean curvature, which was obviously inspired by it. Given $M, \bar{M} \subset \mathbb{R}^3$, with normals ν and $\bar{\nu}$, and an isometry $\phi: M \rightarrow \bar{M}$, we call $p \in M$ a "congruence point" if $\phi^*(\bar{\Pi}(\phi(p))) = \Pi(p)$. If ϕ is not the restriction of a Euclidean motion, then by analyticity the congruence points are isolated. At all other points p , Lemma 13 shows that

$$\det[\phi^*(\bar{\Pi}(\phi(p))) - \Pi(p)] < 0,$$

where the bilinear functions $\Pi(p)$ and $\phi^*(\bar{\Pi}(\phi(p)))$ are regarded as linear transformations on M_p by means of the metric on M_p . It follows that the linear transformation corresponding to the difference $\phi^*(\bar{\Pi}(\phi(p))) - \Pi(p)$ has two eigenspaces, one with a positive eigenvalue, and one with a negative eigenvalue. By picking the one with the positive eigenvalue, say, we obtain a 1-dimensional distribution defined everywhere on M except at the congruence points. At each congruence point we can define the index of the distribution, and the sum of these indices is 2 if M is homeomorphic to S^2 (Theorem 4-20). On the other hand, Cohn-Vossen showed that if M has positive curvature, then the index would have to be negative. The Bibliography will guide the interested reader to descriptions of Cohn-Vossen's argument, as well as alternative arguments and refinements introduced later. We merely mention here that the assumption of analyticity can be dropped by using appropriate results about partial differential equations, and that the whole argument can be formalized to yield an "index method", which has been successfully used in studying certain questions in surface theory; it is one of the few methods which has never yet been generalized to higher dimensions.

We will now return to the use of integral formulas, and prove a result similar to Cohn-Vossen's, which although not strictly speaking a rigidity theorem, nevertheless seems to belong in this chapter since it is the uniqueness part of Minkowski's Problem (page 156). The original proof was based on the general "Brunn-Minkowski inequality" for the "mixed volumes" of convex sets (see Bonnesen-Fenchel [1]). The present proof, obviously inspired by Herglotz' proof of Cohn-Vossen's theorem, is due to Chern.

15. THEOREM (MINKOWSKI). Let M be a compact surface with $K > 0$ everywhere, and let $f, \bar{f}: M \rightarrow \mathbb{R}^3$ be two isometric imbeddings with $N = \bar{N}$.

Then f and \bar{f} differ by a translation. (Alternatively stated, if two compact convex surfaces in \mathbb{R}^3 with everywhere positive curvatures have the same curvatures at points where the normals are parallel, then one surface is a translate of the other.)

PROOF. Since $K > 0$, the maps $N, \bar{N}: M \rightarrow S^2 \subset \mathbb{R}^3$ are diffeomorphisms, so we can consider the imbeddings

$$\begin{aligned} g &= f \circ N^{-1}: S^2 \rightarrow \mathbb{R}^3 \\ \bar{g} &= \bar{f} \circ \bar{N}^{-1}: S^2 \rightarrow \mathbb{R}^3. \end{aligned}$$

These imbeddings have the property that $p \in S^2$ is normal to the tangent plane of $g(S^2)$ at $g(p)$, and similarly for \bar{g} . Thus the normal maps $\xi, \bar{\xi}: S^2 \rightarrow \mathbb{R}^3$ for g and \bar{g} are both the identity map $\text{id}: S^2 \rightarrow S^2$. Since $N = \bar{N}$, and f, \bar{f} are isometries by hypothesis, the maps g, \bar{g} induce the same metric on S^2 . It clearly suffices to show that g and \bar{g} differ by a translation.

Consider the 1-form

$$\omega = (g \times \bar{g}) \cdot d\bar{g}$$

on S^2 . We have

$$\begin{aligned} (1) \quad d\omega &= (dg \times \bar{g}) \cdot d\bar{g} + (g \times d\bar{g}) \cdot d\bar{g} \\ &= -\bar{g} \cdot (dg \times d\bar{g}) + g \cdot (d\bar{g} \times d\bar{g}). \end{aligned}$$

To calculate $d\omega$ explicitly we consider a positively oriented moving frame on S^2 which is orthonormal for the usual metric on S^2 . If we move the vectors $X_i(p)$ over to $g(p)$, then the translated vectors, $Y_i(g(p))$, will be tangent to $g(S^2)$ at $g(p)$, since $\xi(p) = p$. Thus we obtain an orthonormal moving frame Y_1, Y_2 on $g(S^2)$. Let θ^1, θ^2 be the dual forms for Y_1, Y_2 and let ψ_β^α be the connection forms for the moving frame (Y_1, Y_2, ν) on $g(S^2)$, where ν is the unit normal field on $g(S^2)$. Define $\bar{\theta}^i$ and $\bar{\psi}_\beta^\alpha$ similarly, using \bar{g} . If we set

$$\eta^i = g^* \theta^i, \quad \bar{\eta}^i = \bar{g}^* \bar{\theta}^i,$$

then for X tangent to S^2 we have

$$\begin{aligned} (2) \quad dg(X) &= \eta^1(X) \cdot Y_1 + \eta^2(X) \cdot Y_2 = \eta^1(X) \cdot X_1 + \eta^2(X) \cdot X_2, \\ d\bar{g}(X) &= \bar{\eta}^1(X) \cdot X_1 + \bar{\eta}^2(X) \cdot X_2, \end{aligned}$$

where the X_i and Y_i are now regarded as \mathbb{R}^3 -valued functions. It follows that we have

$$\begin{aligned} dg \times d\bar{g} &= (\eta^1 \wedge \bar{\eta}^2 - \eta^2 \wedge \bar{\eta}^1) \cdot \text{id}, \\ d\bar{g} \times d\bar{g} &= 2(\bar{\eta}^1 \wedge \bar{\eta}^2) \cdot \text{id}, \end{aligned}$$

so that (1) becomes

$$(3) \quad d\omega = \bar{h}(\eta^1 \wedge \bar{\eta}^2 - \eta^2 \wedge \bar{\eta}^1) - 2h(\bar{\eta}^1 \wedge \bar{\eta}^2),$$

where h and \bar{h} are the support functions of g and \bar{g} , respectively.

Now we have to relate the forms $\eta^i, \bar{\eta}^i$ to the dual forms for the moving frame X_1, X_2 on S^2 . On $g(S^2)$ we have

$$\psi_i^3 = \sum_j l_{ij} \theta^j$$

where $l_{ij} = \Pi(Y_i, Y_j)$; hence

$$g^* \psi_i^3 = \sum_j (l_{ij} \circ g) \cdot \eta^j.$$

But

$$\begin{aligned} g^* \psi_i^3(X) &= \psi_i^3(g_* X) = -\psi_3^i(g_* X) \\ &= \langle -v_* g_* X, Y_i \rangle \\ &= \langle -d(v \circ g)(X), Y_i \rangle \\ &= \langle -d\xi(X), Y_i \rangle \\ &= \langle -X, Y_i \rangle = \langle -X, X_i \rangle. \end{aligned}$$

So if ζ^1, ζ^2 are the dual forms for X_1, X_2 , so that $\zeta^1 \wedge \zeta^2 = dA$, the volume element of S^2 , then

$$\zeta^i = - \sum_j (l_{ij} \circ g) \cdot \eta^j.$$

If λ is the 2×2 matrix of functions on S^2 defined by

$$\lambda(p) = (l_{ij}(g(p)))^{-1},$$

then

$$\eta^i = - \sum_j \lambda_{ij} \cdot \zeta^j.$$

Similarly,

$$\bar{\eta}^i = - \sum_j \bar{\lambda}_{ij} \zeta^j,$$

where $\bar{\lambda}$ is the inverse of the matrix $(\bar{l}_{ij} \circ \bar{g})$. Since

$$\det(l_{ij} \circ g) = \det(\bar{l}_{ij} \circ \bar{g}) = K,$$

where K is the curvature for the metrics $g^*\langle \ , \ \rangle$ and $\bar{g}^*\langle \ , \ \rangle$ on S^2 , we likewise have

$$\det(\lambda_{ij}) = \det(\bar{\lambda}_{ij}) = K^{-1}.$$

Now we have

$$\bar{\eta}^1 \wedge \bar{\eta}^2 = (\det \bar{\lambda}) \zeta^1 \wedge \zeta^2 = K^{-1} dA$$

and

$$\eta^1 \wedge \bar{\eta}^2 - \eta^2 \wedge \bar{\eta}^1 = (\lambda_{11} \bar{\lambda}_{22} + \bar{\lambda}_{11} \lambda_{22} - 2\lambda_{12} \bar{\lambda}_{12}) dA.$$

Calculating as in the preceding proof, we see that

$$\eta^1 \wedge \bar{\eta}^2 - \eta^2 \wedge \bar{\eta}^1 = [2K^{-1} - \det((d\bar{v})^{-1} - (dv)^{-1})] dA.$$

So equation (3) becomes

$$d\omega = -\{\bar{h} \det((d\bar{v})^{-1} - (dv)^{-1}) + 2K^{-1}(\bar{h} - h)\} dA.$$

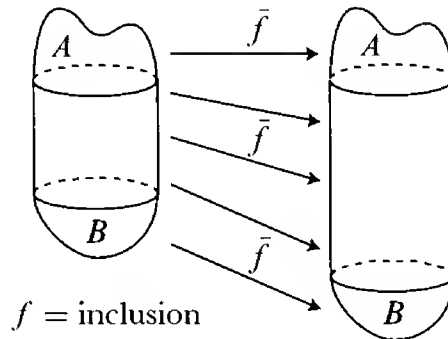
Hence

$$\int_{S^2} \bar{h} \det((d\bar{v})^{-1} - (dv)^{-1}) dA = - \int_{S^2} 2K^{-1}(\bar{h} - h) dA.$$

Then Lemma 13 shows that the left side is ≤ 0 , so that the right side is also ≤ 0 . By symmetry we conclude that the right side is 0, and then by Lemma 13 again that in fact $dv = d\bar{v}$. Thus there is a Euclidean motion A such that $\bar{g} = A \circ g$ and $\bar{v} = A_* v$. The latter implies that A is a translation. ♦

Sometimes this result is expressed in a way that looks quite different: Let $M, \bar{M} \subset \mathbb{R}^3$ be compact surfaces with $K, \bar{K} > 0$ everywhere, and let $\phi: M \rightarrow \bar{M}$ be a map such that $\bar{K}(\phi(p)) = K(p)$ and such that ϕ preserves the third fundamental forms. Then ϕ is the restriction of a Euclidean motion. To see the equivalence of this statement and Theorem 15, just note that by Proposition 2-7, the normal maps of M and \bar{M} are congruent, so after rotating \bar{M} suitably we have two surfaces satisfying the hypothesis of Theorem 15.

Unlike the last few results, Theorem 15 is not true if we allow $K \geq 0$; a counterexample is shown below. Notice, however, that the components of the



set where $K \neq 0$ on one surface do differ from the corresponding components in the other surface by a translation. It is not hard to see that this is always the case, and that Theorem 15 remains true if the set where $K = 0$ is nowhere dense. (For the alternative statement of the theorem surely the only reasonable situation is that in which N is one-one.)

In our next result, the condition $K > 0$ is more crucial, for we are going to consider the sum

$$\frac{1}{k_1} + \frac{1}{k_2} = \frac{2H}{K}$$

of the reciprocals of the principal curvatures; the reciprocals are classically called the “radii of principal curvature”. After constructing a proof of Minkowski’s Theorem by means of integral formulas, Chern then succeeded in constructing a similar proof for the following result, which is actually older than any yet considered.

16. THEOREM (CHRISTOFFEL). Let M be a compact surface with $K > 0$ everywhere, and let $f, \bar{f}: M \rightarrow \mathbb{R}^3$ be two imbeddings with $N = \bar{N}$ such that

$$\frac{1}{k_1} + \frac{1}{k_2} = \frac{1}{\bar{k}_1} + \frac{1}{\bar{k}_2},$$

where the functions k_i and \bar{k}_i on M are the principal curvatures at the corresponding points of $f(M)$ and $\bar{f}(M)$, respectively. Then f and \bar{f} differ by a translation.

PROOF. We introduce the imbeddings $g, \bar{g}: S^2 \rightarrow \mathbb{R}^3$ of the previous proof, and we will use all the notation and formulas from that proof. In addition, we note that for the normal map $\xi = \text{id}$ of g we have, of course,

$$d\xi(X) = X = \zeta^1(X) \cdot X_1 + \zeta^2(X) \cdot X_2,$$

from which it follows that

$$\begin{aligned} (1) \quad d\xi \times dg &= (\zeta^1 \wedge \eta^2 - \zeta^2 \wedge \eta^1) \cdot \xi \\ &= -(\lambda_{22} + \lambda_{11})\xi dA \\ &= -(k_1^{-1} + k_2^{-1})\xi dA. \end{aligned}$$

Similarly, since $\xi = \bar{\xi}$, we have

$$(2) \quad d\xi \times d\bar{g} = -(\bar{k}_1^{-1} + \bar{k}_2^{-1})\xi dA.$$

Consider first the 1-form

$$\omega_1 = (g \times \xi) \cdot d\bar{g}.$$

We have

$$\begin{aligned} d\omega_1 &= (dg \times \xi) \cdot d\bar{g} + (g \times d\xi) \cdot d\bar{g} \\ &= -\xi \cdot (dg \times d\bar{g}) + g \cdot (d\xi \times d\bar{g}) \\ &= -\xi \cdot (dg \times d\bar{g}) - (\bar{k}_1^{-1} + \bar{k}_2^{-1})g \cdot \xi dA \quad \text{by (2)} \\ &= -\xi \cdot (dg \times d\bar{g}) + h(\bar{k}_1^{-1} + \bar{k}_2^{-1}) dA. \end{aligned}$$

Similarly, for the 1-form

$$\omega_2 = (g \times \xi) \cdot dg$$

we have

$$d\omega_2 = -\xi \cdot (dg \times dg) + h(k_1^{-1} + k_2^{-1}) dA.$$

Since $k_1^{-1} + k_2^{-1} = \bar{k}_1^{-1} + \bar{k}_2^{-1}$, we obtain the integral formula

$$(3) \quad \int_{S^2} \xi \cdot (dg \times d\bar{g}) - \xi \cdot (dg \times dg) dA = 0.$$

By interchanging the roles of g and \bar{g} we also obtain

$$(4) \quad \int_{S^2} \xi \cdot (d\bar{g} \times dg) - \xi \cdot (d\bar{g} \times d\bar{g}) dA = 0.$$

Adding (3) and (4) we obtain an integral formula

$$(*) \quad \int_{S^2} I dA = 0.$$

Now we have

$$\begin{aligned} \xi \cdot (dg \times d\bar{g}) - \xi \cdot (dg \times dg) &= (\eta^1 \wedge \bar{\eta}^2 - \eta^2 \wedge \bar{\eta}^1) \xi \cdot \xi - 2(\eta^1 \wedge \eta^2) \xi \cdot \xi \\ &= (\lambda_{11}\bar{\lambda}_{22} + \bar{\lambda}_{11}\lambda_{22} - 2\lambda_{12}\bar{\lambda}_{12} - 2(\lambda_{11}\lambda_{22} - \lambda_{12}^2)) dA, \\ \xi \cdot (d\bar{g} \times dg) - \xi \cdot (d\bar{g} \times d\bar{g}) &= (\bar{\lambda}_{11}\lambda_{22} + \lambda_{11}\bar{\lambda}_{22} - 2\lambda_{12}\bar{\lambda}_{12} - 2(\bar{\lambda}_{11}\bar{\lambda}_{22} - \bar{\lambda}_{12}^2)) dA. \end{aligned}$$

So the integrand I in $(*)$ is

$$\begin{aligned} I &= 2(\lambda_{11}\bar{\lambda}_{22} + \bar{\lambda}_{11}\lambda_{22} - 2\lambda_{12}\bar{\lambda}_{12}) - 2(\lambda_{11}\lambda_{22} - \lambda_{12}^2) - 2(\bar{\lambda}_{11}\bar{\lambda}_{22} - \bar{\lambda}_{12}^2) dA \\ &= -2[(\bar{\lambda}_{11} - \lambda_{11})(\bar{\lambda}_{22} - \lambda_{22}) - (\bar{\lambda}_{12} - \lambda_{12})^2] dA. \end{aligned}$$

Since $\lambda_{11} + \lambda_{22} = \bar{\lambda}_{11} + \bar{\lambda}_{22}$ by hypothesis, we can write

$$\begin{aligned} I &= -2[(\lambda_{22} - \bar{\lambda}_{22})(\bar{\lambda}_{22} - \lambda_{22}) - (\bar{\lambda}_{12} - \lambda_{12})^2] dA \\ &= 2[(\bar{\lambda}_{22} - \lambda_{22})^2 + (\bar{\lambda}_{12} - \lambda_{12})^2] dA. \end{aligned}$$

So the integrand I in $(*)$ is everywhere ≥ 0 . Hence it must be everywhere $= 0$. Hence $\lambda_{ij} = \bar{\lambda}_{ij}$, implying that $dv = d\bar{v}$, and the proof is complete, as before. ♦

In Christoffel's time the radii of principal curvature were regarded as the fundamental entities (which is pretty awkward when $K = 0$), so Theorem 16 was the natural result to try to prove. Nowadays, of course, the result looks rather weird and we would like to formulate it for $H = \frac{1}{2}(k_1 + k_2)$ instead. Oddly enough, this more reasonable looking problem hasn't been solved. The pair of surfaces pictured on page 203 give a counterexample of sorts, but I do not know of any counterexample which is strictly convex. This same pair of surfaces illustrates the need for the final hypothesis appearing in the following result along these lines, which replaces the conditions on the normal maps by one on the imbeddings themselves (as some sort of compensation for the stringency of this hypothesis, notice that no hypothesis on K is required).

17. THEOREM (HOPF AND VOSS). Let M be a compact surface, and let $f, \bar{f}: M \rightarrow \mathbb{R}^3$ be two imbeddings with $H = \bar{H}$ such that $\bar{f}(p) - f(p)$ is always parallel to a fixed vector $v \in \mathbb{R}^3$. Suppose, moreover, that $f(M)$ and $\bar{f}(M)$ do not contain a portion of a cylinder with generators parallel to v . Then f and \bar{f} differ by a translation in the direction of v .

PROOF. Write

$$\bar{f} = f + \alpha \cdot v$$

for some function α on M . For any $X_1, X_2 \in M_p$ we have

$$\begin{aligned} d\bar{f}(X_1) \times d\bar{f}(X_2) &= [df(X_1) + d\alpha(X_1) \cdot v] \times [df(X_2) + d\alpha(X_2) \cdot v] \\ &= df(X_1) \times df(X_2) \\ &\quad + [d\alpha(X_1) \cdot v \times df(X_2) - d\alpha(X_2) \cdot v \times df(X_1)]. \end{aligned}$$

So

$$d\bar{f} \times d\bar{f} = df \times df + 2(d\alpha \cdot v \times df).$$

By (I), this is equivalent to

$$(1) \quad \bar{N} d\bar{A} = N dA + (d\alpha \cdot v \times df).$$

Consider the 1-form

$$\omega_1 = (\alpha \cdot v \times n) \cdot df.$$

We have

$$\begin{aligned} (2) \quad d\omega_1 &= (d\alpha \cdot v \times N) \cdot df + (\alpha \cdot v \times dN) \cdot df \\ &= -N \cdot (d\alpha \cdot v \times df) + \alpha \cdot v \cdot (dN \times df) \\ &= dA - (N \cdot \bar{N}) d\bar{A} + 2\alpha H(v \cdot N) dA \quad \text{by (1) and (I).} \end{aligned}$$

Similarly, for the 1-form

$$\omega_2 = (\alpha \cdot v \times \bar{N}) \cdot df,$$

we have

$$\begin{aligned}
 (3) \quad d\omega_2 &= (d\alpha \cdot v \times \bar{N}) \cdot df + (\alpha \cdot v \times d\bar{N}) \cdot df \\
 &= (d\alpha \cdot v \times \bar{N}) \cdot df + (\alpha \cdot v \times d\bar{N}) \cdot d\bar{f}, \\
 &\quad \text{since } d\bar{f} = df + d\alpha \cdot v \\
 &= -\bar{N} \cdot (d\alpha \cdot v \times df) + \alpha \cdot v \cdot (d\bar{N} \times d\bar{f}) \\
 &= N \cdot \bar{N} dA - d\bar{A} + 2\alpha \bar{H}(v \cdot \bar{N}) d\bar{A} && \text{by (I) and (I)} \\
 &= N \cdot \bar{N} dA - d\bar{A} + 2\alpha \bar{H}(v \cdot N) dA && \text{by (I).}
 \end{aligned}$$

From (2) and (3) we derive the integral formula

$$2 \int_M \alpha (\bar{H} - H)(v \cdot N) dA = \int_M (1 - N \cdot \bar{N})(dA + d\bar{A}).$$

Since $H = \bar{H}$, we thus have

$$\int_M (1 - N \cdot \bar{N})(dA + d\bar{A}) = 0.$$

But $N \cdot \bar{N} \leq 1$, so we must have $N \cdot \bar{N} = 1$ everywhere, and hence $N = \bar{N}$ everywhere. We will use this to show that α is constant.

For simplicity assume that v points along the z -axis. The final hypothesis implies that $N(p)$ has non-zero z -component for all points p in a dense open set. If p is such a point, then $f(M)$ and $\bar{f}(M)$ can be represented near p as the graphs of two functions, g and \bar{g} , say. Then the normals are given by

$$\frac{(g_1, g_2, -1)}{\sqrt{1 + g_1^2 + g_2^2}} \quad \text{and} \quad \frac{(\bar{g}_1, \bar{g}_2, -1)}{\sqrt{1 + \bar{g}_1^2 + \bar{g}_2^2}}.$$

Since these normals are everywhere equal, we must have $g_i = \bar{g}_i$, so α is constant in a neighborhood of p . ♦

Before returning to rigidity theory proper, we will take this opportunity to mention that Minkowski's Theorem and Christoffel's Theorem have generalizations to compact hypersurfaces $M \subset \mathbb{R}^{n+1}$ with all principal curvatures

$k_1, \dots, k_n > 0$. For each $i = 1, \dots, n$ we can consider the elementary symmetric polynomial

$$P_i = \sigma_i \left(\frac{1}{k_1}, \dots, \frac{1}{k_n} \right).$$

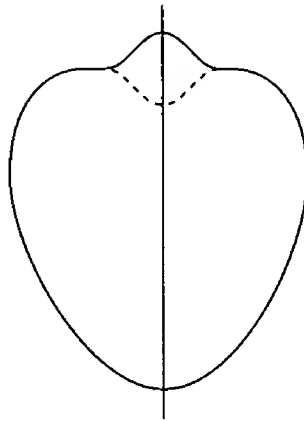
If M and \bar{M} are two such hypersurfaces, and for some i the functions P_i and \bar{P}_i agree at points of M and \bar{M} where the normals are parallel, then one hypersurface is a translate of the other. A proof may be found in Chern [1]. (Our proofs of Theorems 15 and 16 are taken from that paper, and are pretty representative of the sort of argument which is used. In fact, the proof of Theorem 15 is a special case of Chern's proof for all $i > 1$, while the proof of Theorem 16 is a special case of Chern's proof for $i = 1$, which needs a separate argument. Chern remarks that that distinction is significant, since the case $i = 1$ involves linear partial differential equations, while the case $i > 1$ involves non-linear ones.)

Another remark is necessary to put Theorem 16 on an equal footing with Theorems 12 and 15. The latter two results give the uniqueness of imbeddings whose existence was discussed in Chapter 11 (Weyl's Problem and Minkowski's Problem). The corresponding existence result for Christoffel's Problem is much more involved, for there are complicated relations which must be satisfied by a given function on S^2 in order for it to be $1/k_1 + 1/k_2$ for some imbedded surface. Many incomplete treatments of this problem have been given, and the correct necessary and sufficient conditions (in all dimensions) were discovered only in 1967 by Firey [1]. One could also seek the conditions on a function in order that it be P_i ($1 < i < n$) for some convex hypersurface, but this problem is perhaps hopelessly complicated.

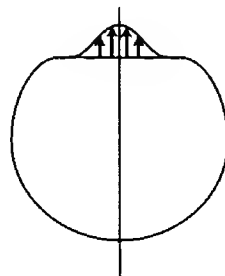
Finally, we want to point out that the higher dimensional generalization of the Minkowski and Christoffel theorems can also be expressed as follows: Let $M, \bar{M} \subset \mathbb{R}^{n+1}$ be compact hypersurfaces with all principal curvatures > 0 , and let $\phi: M \rightarrow \bar{M}$ be a map such that $\bar{P}_i(\phi(p)) = P_i(p)$ and such that ϕ preserves the third fundamental forms. Then ϕ is the restriction of a Euclidean motion. The argument is just the same as on page 203, using Problem 7-18 in place of Proposition 2-7. In this connection it is interesting that É. Cartan [2] raised the possibility of studying surfaces by means of their *second* fundamental form, rather than their first. For example, he showed that if \mathbf{II} is positive definite, then the curvature of (M, \mathbf{II}) can be written in terms of the ordinary principal curvatures and their derivatives. Grove [1] used integral formulas to prove that a diffeomorphism of compact convex surfaces which preserves \mathbf{II} and the Gaussian curvature $K = k_1 k_2$ is the restriction of a Euclidean motion, and Walden [1] used the index method to prove the same result if either $k_1 + k_2$ or $k_1^{-1} + k_2^{-1}$ is

preserved (as well as if $k_1^2 + k_2^2$ or $k_1^{-2} + k_2^{-2}$ is preserved). On the other hand, it is perfectly conceivable that a diffeomorphism of compact convex surfaces is the restriction of a Euclidean motion if it merely preserves \mathbf{II} , and perhaps the surfaces need not even be convex. In higher dimensions, the only result is that of Gardner [1], proved using integral formulas, which generalizes Grove's result to the case where \mathbf{II} and $k_1 \cdots k_n$ are preserved.

All our rigidity results have required the hypothesis of convexity, so it is only natural to wonder what happens in the case of non-convex surfaces. There is a standard example, illustrated below, of two compact rotation surfaces which

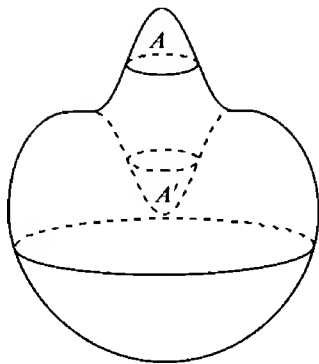


are isometric but not congruent. To be sure, this example is rather unsatisfying, since it is merely C^∞ , and cannot be modified to be analytic. Moreover, the surface consists of two parts which are individually kept rigid, but which are glued together in two different ways along a plane curve where $K = 0$. Finally, this example is merely a reflection of the fact that the plane has infinitesimal bendings which vanish outside a compact set. It could have been obtained by starting with a surface containing a portion of a plane, finding an infinitesimal



bending Z which vanishes outside the planar region, and then mapping $p + Z(p)$ to $p - Z(p)$, which is an isometry, by Lemma 7, but not the restriction of a Euclidean motion, by Lemma 8. We can at least show that there is no bending connecting our two non-congruent isometric C^∞ surfaces. In fact, there is not

even a bending taking the small region A of positive curvature pictured below to its corresponding region A' in the other surface, the region A' being simply



the mirror image of A . To prove this, we simply recall that any surface of positive curvature has a natural orientation, which has to be preserved during the bending.

Long before such C^∞ trickiness was in vogue, Cohn-Vossen [1] had investigated infinitesimally bendable rotation surfaces by quite different methods, and he found C^∞ rotation surfaces with non-trivial C^2 infinitesimal bendings. Afterwards, Rembs [1] managed to obtain an example where the rotation surface is analytic; but the infinitesimal bending is still only C^2 (a point which is by no means made clear in the paper). Applying Lemmas 7 and 8 in this case we merely obtain two C^2 isometric non-congruent surfaces, which seems a lot worse than C^∞ ; but at least this example is not a bald-faced trick like the previous one. Since the example is interesting, but nevertheless rather disappointing, it has been relegated to an Addendum.

It seems that at present it is simply unknown whether every analytic compact surface is unwarpable. *A fortiori* it is unknown whether every analytic compact surface is unbendable; it certainly seems likely that even C^∞ compact surfaces are unbendable.

There is only one crumb of comfort which we can offer in this dismal situation. It is known that a torus of revolution is unwarpable. Such a torus M has



the property that

$$\int_{M^+} K \, dA = 4\pi,$$

where $M^+ = \{p \in M : K(p) > 0\}$; the closure $C(M^+)$ of M^+ is a compact surface with boundary such that $\partial C(M^+)$ is the union of plane curves along which the tangent space of M is constant. Suppose that $\phi: M \rightarrow \bar{M}$ is an isometry. Then also

$$\int_{\bar{M}^+} K dA = 4\pi,$$

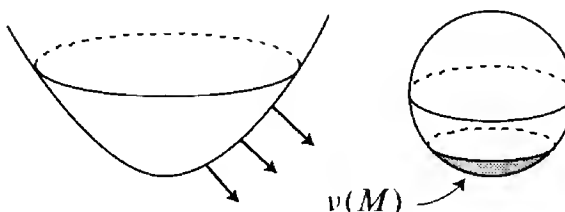
so it follows from Theorem 6-16 that $C(\bar{M}^+)$ is a compact surface of the same sort as $C(M^+)$. Now apply Lemma 14 to $f =$ inclusion map of $C(M^+)$ and $\tilde{f} = \phi: C(M^+) \rightarrow C(\bar{M}^+)$. The terms $\tau_g, \kappa_n, \bar{\tau}_g, \bar{\kappa}_n$ are all 0, since the normal is constant along each component of $\partial C(M^+)$ and $\partial C(\bar{M}^+)$. So we have simply

$$0 = \int_{C(M^+)} h \det(d\bar{N} - dN) dA + 2 \int_{C(M^+)} \bar{H} - H dA,$$

the same equality which we used in the proof of Theorem 12. Then the same argument which was used in this proof shows that $\phi: C(M^+) \rightarrow C(\bar{M}^+)$ is the restriction of a Euclidean motion.

This already shows that any *analytic* surface of minimal total absolute curvature is unwarpageable in the class of *analytic* surfaces, a result originally due to Alexandrov [4]. A proof that ϕ is also the restriction of a Euclidean motion on the part of the surface with $K < 0$ has been given by Nirenberg [2]. The proof, which involves a discussion of hyperbolic equations, requires some additional, rather unsatisfactory, hypotheses, but these hypotheses are satisfied at least in the special case of a torus of revolution.

We can also ask about the rigidity of complete convex non-compact surfaces. It is not hard to see (compare the pictures on pg. IV.84) that the normal map v of such a surface M always lies in a hemisphere, so that

$$\int_M K dA \leq 2\pi.$$


The first result on complete convex non-compact surfaces was the surprising theorem of Olowjanischnikow [1] that M is *warpageable* if $\int_M K dA < \pi$. Olowjanischnikow's proof uses the methods of the Russian school of differential

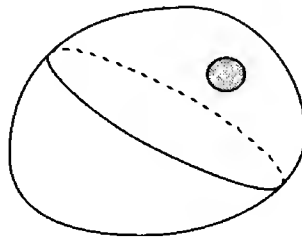
geometry, which was briefly discussed in Chapter 11. As we have already mentioned, these methods, though intricate and difficult, allow one to prove certain results for surfaces which are merely continuous. For example, Pogorelov {1} has proved Cohn-Vossen's theorem for arbitrary convex surfaces: if M and \bar{M} are the boundaries of compact convex sets (with non-empty interiors) in \mathbb{R}^3 , and $\phi: M \rightarrow \bar{M}$ is a map which preserves lengths of curves, then ϕ is a congruence. Similarly, Olowjanischnikow's result holds whenever M is the boundary of a closed non-compact convex set (with non-empty interior) in \mathbb{R}^3 . Pogorelov {2; pg. 114} also showed that any surface isometric to such a surface M may be joined to M by a continuous bending. To my knowledge, no one has ever provided simpler proofs of these results when the surfaces considered are C^∞ .

When our complete convex surface M has

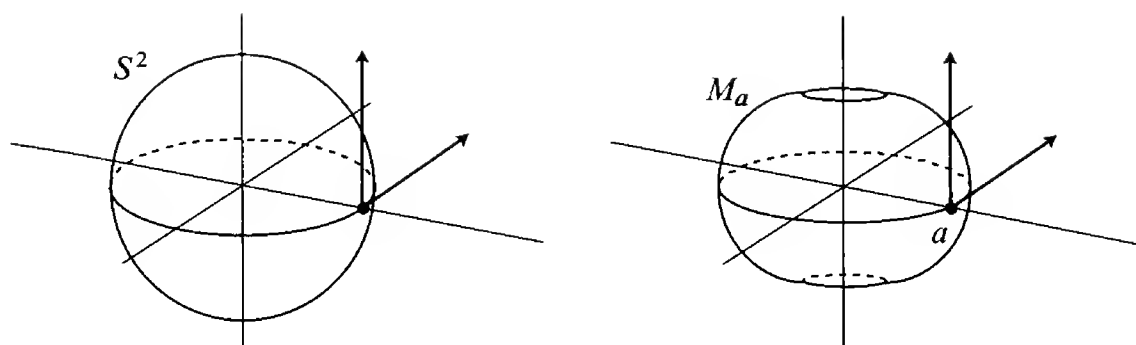
$$\int_M K dA = 2\pi,$$

it is unwarpable. The proof of this is due to Pogorelov [2]. We have already seen that such surfaces, although unwarpable, and hence unbendable, may nevertheless be infinitesimally bendable; and I think that this is the only known instance of such a phenomenon.

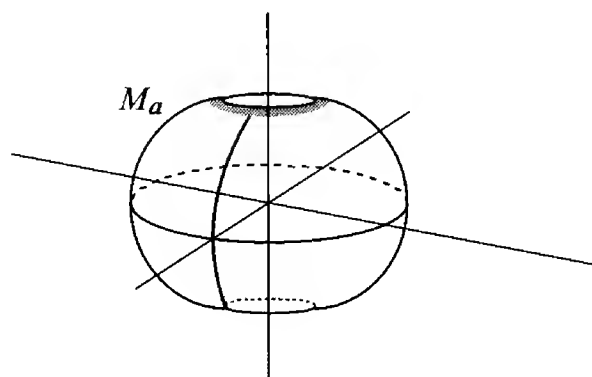
As opposed to the complete convex surfaces, consider what happens when we delete a set with non-empty interior from a convex surface. Is it bendable,



or warpable, or infinitesimally bendable? The promptings of intuition seem to vary from person to person, and historically there was considerable confusion on the question. We claim first of all that any open set $U \subset S^2$ whose closure \bar{U} is contained in an open hemisphere of S^2 is warpable. For this purpose we consider the rotation surface M_a of constant curvature 1 given on pg. III.163, with $a > 1$. Locally there is an isometry $S^2 \rightarrow M_a$ taking the tangent vectors $(0, 1, 0)$ and $(0, 0, 1)$ at $(1, 0, 0) \in S^2$ to the tangent vectors $(0, 1, 0)$ and $(0, 0, 1)$ at $(a, 0, 0) \in M_a$. If a is sufficiently close to 1, then this isometry can be extended to cover $\bar{U} \subset$ the hemisphere $\{p \in S^2 : p^1 > 0\}$. The image of U is contained



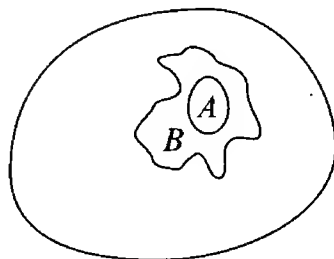
in the open set $V \subset M_a$ which is obtained by deleting the profile curve of M_a in the left half (x, z) -plane, as well as a neighborhood of the top and bottom boundary curves. We also claim that this open set V is bendable. To prove this



we consider all the rotation surfaces $M_{a'}$ for a' close to a . Then there will be an isometry $f_{a'}: V \rightarrow M_{a'}$ which takes the tangent vectors $(0, 1, 0)$ and $(0, 0, 1)$ at $(a, 0, 0) \in M_a$ to the tangent vectors $(0, 1, 0)$ and $(0, 0, 1)$ at $(a', 0, 0) \in M_{a'}$. The 1-parameter family of isometries $\{f_{a'}\}$ gives us the bending.

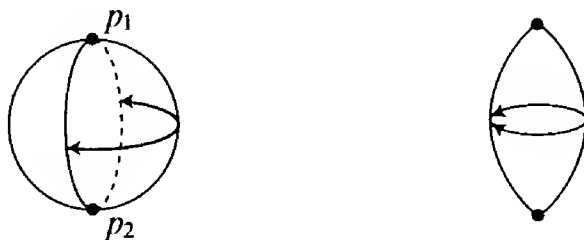
The warpability of U was noted in a paper by Liebmann [1] in 1900, which also offered up a proof that U is not warpable if it contains a closed hemisphere. Fifteen years later, Blaschke [1] observed that the proof, “as simple as this assertion may appear”, was incorrect. In 1919 Liebmann [2] showed that in fact the sphere minus any disc, no matter how small, is bendable. He did this by specifically constructing the bending, using other classical examples of open surfaces of constant curvature 1. A physically intuitive argument, involving soap bubbles, is given in Hilbert and Cohn-Vossen [1]. Liebmann was then willing to conjecture that any convex surface with $K > 0$ everywhere is bendable after a small disc is removed. The infinitesimal bendability of such surfaces was proved by Cohn-Vossen [2] in 1927, and bendability was proven by Hellwig [1] in 1955; both proofs require facts about partial differential equations, with more difficult theorems required for bendability. (This question has also been treated by the Russian school; see Pogorelov [2; pg. 104].)

It should be noted that the bendability of a convex surface M with a set $A \subset M$ removed does not necessarily imply the bendability of $M - B$ for $A \subset B$. For conceivably the bending of $M - A$ might always be constant on



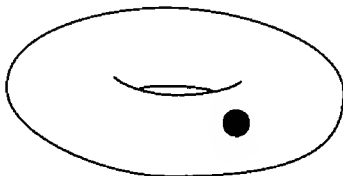
$M - B$! But at least we do not have to worry about this anomaly if M is analytic, with $K > 0$ everywhere. For it then follows from the results of Chapter 11 that all warpings of $M - A$ are also analytic.

We can also ask what happens when we delete even smaller sets from a convex surface M . Hilbert and Cohn-Vossen [1] claim that the sphere minus any segment of a great circle is bendable, but I have never seen a reference to such a result. I am almost certain that nothing similar is known when the sphere is replaced by an arbitrary convex surface M . As recently as 1971, Green and Wu [1] showed that if only finitely many points are removed from a compact surface M with $K \geq 0$ everywhere, then the resulting surface $M' = M - \{p_1, \dots, p_k\}$ is unwarpable. On the other hand, Pogorelov [1] had already shown that if $k \geq 2$, then M' is warpable as an immersion: there is an isometric immersion of M' which is not an imbedding (in fact, there are infinitely many inequivalent isometric immersions). In one special case this is easy to see: Take M' to be $S^2 - \{p_1, p_2\}$, where p_1 and p_2 are the north and south poles, and consider the surface of revolution M_a on pg. III.163 with $a = 1/2$; it is not hard to compute



that the area of M_a is just $1/2$ the area of S^2 . Then there will be an isometry of a closed hemisphere of $S^2 - \{p_1, p_2\}$ onto M_a which takes the two semi-circles on the boundary onto the same curve in M_a , namely its profile curve in the left half (x, z) -plane. By using a similar map on the other hemisphere of $S^2 - \{p_1, p_2\}$ we obtain a local isometry $S^2 - \{p_1, p_2\} \rightarrow M_a$ which is a double covering.

To end this discussion of holey surfaces we mention one more indication of the abysmal state of our ignorance: It is not known whether the standard torus minus a disc is bendable, or even warpable.



We now turn our attention to purely local results about rigidity, where further surprises are in store for us. We will begin by examining some of the classical results along this line, partly to give an idea of the sort of questions which used to be investigated, and partly because some of these questions throw great light on the geometric aspects of the Darboux equation.

Consider a surface $M \subset \mathbb{R}^3$ and an arclength parameterized curve $\tilde{c}: [a, b] \rightarrow M$. Given another arclength parameterized curve $c: [a, b] \rightarrow \mathbb{R}^3$, we ask whether there is a neighborhood \tilde{V} of $\tilde{c}([a, b])$ in M and an isometry $\phi: \tilde{V} \rightarrow V$ of \tilde{V} onto a surface $V \subset \mathbb{R}^3$ such that $\phi \circ \tilde{c} = c$. In other words, we want to know to what extent a curve \tilde{c} on M can be changed in a local warping. We might as well assume that M is the image of an isometric immersion $\tilde{f}: (U, (g_{ij})) \rightarrow \mathbb{R}^3$, where $U \subset \mathbb{R}^2$ is an open set containing $[a, b] \times \{0\}$, and $\tilde{c}(x) = \tilde{f}(x, 0)$ for $x \in [a, b]$. On $[a, b] \times \{0\}$ we can compute the geodesic curvature $\tilde{\kappa}_g$ of \tilde{c} . If the isometry ϕ exists, then the geodesic curvature κ_g of c on V must be the same as $\tilde{\kappa}_g$. On the other hand, if c is a curve on any surface V whatsoever, then its geodesic curvature κ_g on V always has absolute value less than or equal to its curvature κ (which is a known function). Thus we see that the isometry ϕ cannot exist unless

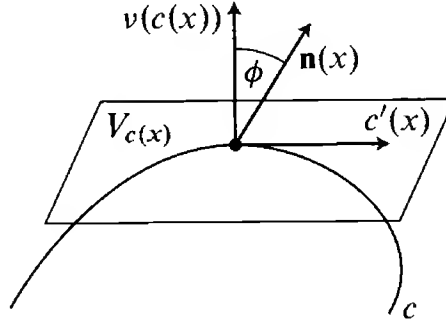
$$|\tilde{\kappa}_g| \leq \kappa.$$

Consider the case where we have the strict inequality $|\tilde{\kappa}_g(x)| < \kappa(x)$ for all x . If the surface V exists, and $\nu(c(x))$ is its normal at $c(x)$, then equation (9) on pg. III.190 shows that we must have

$$\tilde{\kappa}_g = \kappa_g = \kappa \cdot \sin \phi,$$

where ϕ is the angle from the principal normal $\mathbf{n}(x)$ of c to $\nu(c(x))$. This equation determines two possible choices for ϕ , and hence two possible choices

for the tangent space $V_{c(x)}$. Consider either of the two possible continuous choices of $V_{c(x)}$ along c . We claim that we can find V with this choice of $V_{c(x)}$.



We will look for V as the image of an isometric immersion $f: U \rightarrow \mathbb{R}^3$ with $c(x) = f(x, 0)$. By changing our coordinate system on U , we can assume that

$$g_{12} = 0, \quad g_{22} = 1; \quad \text{and} \quad g_{11} = 1 \quad \text{along } [a, b] \times \{0\}.$$

We easily compute that in this case the curve $x \mapsto (x, 0)$ has geodesic curvature

$$(1) \quad \tilde{\kappa}_g(x) = -\frac{1}{2}(g_{11})_y.$$

Now the vector $\mathbf{v} = (0, 1)_{(x, 0)} \in U_{(x, 0)}$ is a unit vector perpendicular to $(1, 0)_{(x, 0)} \in U_{(x, 0)}$. Obviously we want the vector

$$f_*(\mathbf{v}) = \chi(x), \quad \text{say}$$

to be a unit vector in the known vector space $V_{c(x)}$, perpendicular to the known tangent vector $c'(x)$. Thus $\chi(x)$ is determined along $[a, b] \times \{0\}$. For the values of $\chi(x)$ so determined we have

$$(2) \quad \begin{cases} \langle \chi(x), c'(x) \rangle = 0 \\ \langle \chi(x), \chi(x) \rangle = 1 \\ \langle \chi(x), c''(x) \rangle = \kappa(x) \cdot \langle \chi(x), \mathbf{n}(x) \rangle \\ \quad = \kappa(x) \cdot \sin \phi \\ \quad = \tilde{\kappa}_g(x) = -\frac{1}{2}(g_{11})_y, \quad \text{by (1).} \end{cases}$$

Now we can use the arguments in the proof of Theorem 11-9. We have the map f defined on $[a, b] \times \{0\}$ by $f(x, 0) = c(x)$, and we want to extend f to a neighborhood of $[a, b] \times \{0\}$ in U . The submanifold $[a, b] \times \{0\}$ isn't necessarily of the form $\exp_0(V_1)$, but that is irrelevant, because we have already determined χ

satisfying (2), which are just the equations (2') on page 151. Moreover, χ is linearly independent of f_x, f_{xx} on $[a, b] \times \{0\}$, as required. Thus we can solve equation (*) in the proof of Theorem 11-9 with the initial conditions $f(x, 0) = c(x)$ and $q(x, 0) = \chi(x)$.

This argument requires that the original surface M and the curves \tilde{c}, c be analytic, in order to apply the Cauchy-Kowalewski theorem. But analyticity is not needed if M has curvature $K < 0$, for in this case the system involved is hyperbolic, as we mentioned on page 154.

For other purposes, it is also important that we examine the classical treatment of this problem, by means of the Darboux equation. We continue to use the special coordinate system on U , with $g_{12} = 0$ and $g_{22} = 1$; and $g_{11} = 1$ along $[a, b] \times \{0\}$. We would like to find $f = (u, v, w)$ by using the three components of the equations

$$\begin{aligned} f(x, 0) &= c(x) \\ f_y(x, 0) &= \chi(x) \end{aligned}$$

as the initial conditions for the solutions u, v, w of the Darboux equation. Unfortunately that simple procedure won't work, since, as we have already seen, there is not that much arbitrariness permitted in the choice of u, v, w . What we have to do is first find a solution w of the Darboux equation with

$$(3) \quad w(x, 0) = c^3(x), \quad w_y(x, 0) = \chi^3(x),$$

and then choose u and v so that we at least have

$$(4) \quad \begin{aligned} u(0, 0) &= c^1(0), & u_x(0, 0) &= c^{1'}(0), & u_y(0, 0) &= \chi^1(0) \\ v(0, 0) &= c^2(0), & v_x(0, 0) &= c^{2'}(0), & v_y(0, 0) &= \chi^2(0), \end{aligned}$$

and so that $f = (u, v, w)$ is an isometry, and consequently satisfies

$$(5) \quad \left\{ \begin{array}{l} g_{11} = \langle f_x, f_x \rangle = u_x \cdot u_x + v_x \cdot v_x + w_x \cdot w_x \\ 0 = \langle f_x, f_y \rangle = u_x \cdot u_y + v_x \cdot v_y + w_x \cdot w_y \\ 1 = \langle f_y, f_y \rangle = u_y \cdot u_y + v_y \cdot v_y + w_y \cdot w_y \\ \quad \quad \quad \Downarrow \text{(as in the proof of Theorem 11-9)} \\ -\frac{1}{2}(g_{11})_y = \langle f_{xx}, f_y \rangle = u_{xx} \cdot u_y + v_{xx} \cdot v_y + w_{xx} \cdot w_y. \end{array} \right.$$

We claim that u and v will automatically have the desired initial conditions on $[a, b] \times \{0\}$. To see this, consider the \mathbb{R}^2 -valued functions

$$\begin{aligned} \alpha(x) &= (c^{1'}(x), c^{2'}(x)) \\ \beta(x) &= (\chi^1(x), \chi^2(x)). \end{aligned}$$

By substituting equations (3) into equations (2), we find that α and β satisfy

$$\begin{aligned} \text{(i)} \quad & \langle \alpha, \beta \rangle = -w_y(x, 0) \cdot w_x(x, 0) \\ \text{(ii)} \quad & \langle \beta, \beta \rangle = 1 - w_y(x, 0) \cdot w_y(x, 0) \\ \text{(iii)} \quad & \langle \alpha', \beta \rangle = -\frac{1}{2}(g_{11})_y - w_y(x, 0) \cdot w_{xx}(x, 0), \end{aligned}$$

while we also have

$$\text{(iv)} \quad \langle \alpha, \alpha \rangle = 1 - w_x(x, 0) \cdot w_x(x, 0) = g_{11}(x, 0) - w_x(x, 0) \cdot w_x(x, 0).$$

Simple arguments (Problem 1) show that the solution of the system of equations (i)–(iv) is completely determined once $\alpha(0), \beta(0)$ are known. But equations (5) show that

$$\begin{aligned} & (u_x(x, 0), v_x(x, 0)) \\ & (u_y(x, 0), v_y(x, 0)) \end{aligned}$$

satisfy this system, while equations (4) insure that

$$\begin{aligned} \alpha(0) &= (u_x(0, 0), v_x(0, 0)) \\ \beta(0) &= (u_y(0, 0), v_y(0, 0)). \end{aligned}$$

It follows that for all $x \in [a, b]$ we have

$$\begin{aligned} (c^1(x), c^2(x)) &= \alpha(x) = (u_x(x, 0), v_x(x, 0)) \\ &\implies (c^1(x), c^2(x)) = (u(x, 0), v(x, 0)) \quad \text{by (4)} \\ (\chi^1(x), \chi^2(x)) &= \beta(x) = (u_y(x, 0), v_y(x, 0)). \end{aligned}$$

Thus u and v will indeed have the initial conditions which we would like, and $f = (u, v, w)$ will be the required isometric imbedding.

We have already noted in the previous chapter that the Darboux equation is hyperbolic when $K < 0$, so we do not need the Cauchy-Kowalewski theorem in that case. In this case there is still one problem remaining, however, for in order to solve the Darboux equations for w with the initial conditions (3), we need to know that the interval $[a, b]$ of the x -axis is free for these initial conditions. Here is the place where the geometry links up beautifully with the analysis. To begin with, suppose we have an isometric immersion $f = (u, v, w): (U, (g_{ij})) \rightarrow \mathbb{R}^3$, and a curve γ in U . Then w is a solution of the Darboux equation, and we

would like to know when γ is free for the initial conditions which we obtain by restricting w to γ . Recall that for the second order PDE

$$F(x, y, u, p, q, r, s, t) = 0,$$

this means (c.f. (I-4) on page 84) that

$$\frac{\partial F}{\partial r} \cdot (v^1)^2 + \frac{\partial F}{\partial s} \cdot v^1 v^2 + \frac{\partial F}{\partial t} \cdot (v^2)^2 \neq 0 \quad \text{at } \gamma(t),$$

where (v^1, v^2) is the normal to γ at t . Since (v^1, v^2) is proportional to $(-\gamma_2', \gamma_1')$, this means that

$$\frac{\partial F}{\partial r} \cdot (\gamma_2')^2 - \frac{\partial F}{\partial s} \cdot \gamma_1' \gamma_2' + \frac{\partial F}{\partial t} \cdot (\gamma_1')^2 \neq 0.$$

Consider the Darboux equation in the form (*) on page 143. Our condition becomes

$$(6) \quad (w_{22} - \Gamma_{22}^1 w_1 - \Gamma_{22}^2 w_2)(\gamma_2')^2 + 2(w_{12} - \Gamma_{12}^1 w_1 - \Gamma_{12}^2 w_2)\gamma_1' \gamma_2' \\ + (w_{11} - \Gamma_{11}^1 w_1 - \Gamma_{11}^2 w_2)(\gamma_1')^2 \neq 0,$$

which by equation (1) on page 143 becomes

$$0 \neq N^3[l_{22}(\gamma_2')^2 + 2l_{12}\gamma_1'\gamma_2' + l_{11}(\gamma_1')^2] \\ = N^3\Pi_f(\gamma', \gamma').$$

Thus the curve γ is free for the initial conditions determined by w if and only if the curve $f \circ \gamma$ on $f(U)$ is nowhere asymptotic, and the tangent plane for $f(U)$ is nowhere vertical along γ . On the other hand, γ is characteristic for these initial conditions if and only if $f \circ \gamma$ is an asymptotic curve, except perhaps at points where the tangent plane of $f(U)$ is vertical. These conditions have the paradoxical feature customarily associated with the Darboux equation: the condition on a single component w of f is stated in terms of the whole map f . When we are merely given initial conditions along a curve, rather than a solution, we simply write out equation (6) as stated. In the situation we are considering, our initial curve is just the interval $[a, b]$ of the x -axis, so that $\gamma_1' = 1$ and

$\gamma_2' = 0$, and we compute that in our special coordinate system on U we have

$$\left. \begin{aligned} [11, 2] &= -\frac{1}{2}(g_{11})_y \\ [12, 1] &= [21, 1] = \frac{1}{2}(g_{11})_y \\ \text{other } [ij, k] &= 0 \end{aligned} \right\} \quad \text{along the } x\text{-axis,}$$

and then

$$\left. \begin{aligned} \Gamma_{12}^1 &= \Gamma_{21}^1 = \frac{1}{2}(g_{11})_y \\ \Gamma_{11}^2 &= -\frac{1}{2}(g_{11})_y \\ \text{other } \Gamma_{ij}^k &= 0 \end{aligned} \right\} \quad \text{along the } x\text{-axis.}$$

Then equation (6) becomes

$$(6') \quad w_{11} + \frac{1}{2}(g_{11})_y w_2 \neq 0, \quad \text{or} \quad (c^3)'' - \tilde{\kappa}_g \chi^3 \neq 0.$$

Now in our situation, χ is not a multiple of c'' . Simply by rotating everything, we can then insure that $(c^3)'' - \tilde{\kappa}_g \chi^3 \neq 0$ on $[a, b] \times \{0\}$. Thus we really can solve the Darboux equation and obtain an isometry $\phi: \tilde{V} \rightarrow V$ with $\phi \circ \tilde{c} = c$ (suitably rotated). Naturally we can then obtain a new isometry ϕ' with $\phi' \circ \tilde{c} = c$.

Things work out quite differently when we try to find an isometry $\phi: \tilde{V} \rightarrow V$ with $\phi \circ \tilde{c} = c$ in the case where $\tilde{\kappa}_g(x) = \kappa(x)$ for all x . If ϕ exists, then we must have $\kappa_g = \tilde{\kappa}_g = \kappa$, so c must be an asymptotic curve on V , which means that V must have $K \leq 0$ along c so M must have $K \leq 0$ along \tilde{c} . We will actually assume $K < 0$, so that the Darboux equations are hyperbolic. We first suppose that $\kappa(x) > 0$ for all x . Then the Beltrami-Enneper Theorem (Theorem 4-7) shows that the torsion τ of c must satisfy

$$\tau(x) = \pm \sqrt{-K(c(x))} = \pm \sqrt{-K(\tilde{c}(x))}.$$

Thus there is, up to Euclidean motions, only one possibility for c . If we are given this curve c , and $\phi: \tilde{V} \rightarrow V$ exists, then $V_{c(x)}$ must be the osculating plane of c at x , so our choice for $\chi(x) = f_y(x, 0)$ must be the principal normal $\mathbf{n}(x)$ of c at $\chi(x)$. In this situation we have $c'' = \kappa \mathbf{n} = \kappa \chi = \tilde{\kappa}_g \chi$, so equation (6') is *not* true; our initial curve is characteristic for the initial conditions. Fortunately, we have complete information about this situation, since we are dealing with a Monge-Ampère equation. Along the x -axis our Darboux equation (from page 143) is

$$[w_{11} + \frac{1}{2}(g_{11})_y w_2] \cdot w_{22} - [w_{12} - \frac{1}{2}(g_{11})_y w_1]^2 = K(1 - w_1^2 - w_2^2).$$

As we pointed out at the end of section 8 of Chapter 10, there is no hope of solving for w_{22} along the x -axis unless we also have

$$(7) \quad -[w_{12} - \tfrac{1}{2}(g_{11})_y w_1]^2 = K(1 - w_1^2 - w_2^2)$$

along the x -axis. Moreover, if this equation does hold, then we can choose w_{22} arbitrarily, and there will be a solution with these initial conditions. We claim that equation (7) holds as a consequence of our choice of c . For,

$$\begin{aligned} [w_{12} - \tfrac{1}{2}(g_{11})_y w_1]^2 &= \{(\mathbf{n}' + \kappa c')^3\}^2 && \text{[the 3 denotes third component]} \\ &= \{(-\kappa c' + \tau \mathbf{b} + \kappa c')^3\}^2 && \text{by Serret-Frenet} \\ &= \{(\tau \mathbf{b})^3\}^2 = \tau^2 (\mathbf{b}^3)^2 \\ &= -K(\mathbf{b}^3)^2, \end{aligned}$$

so we just have to show that

$$(\mathbf{b}^3)^2 = 1 - w_1^2 - w_2^2 = 1 - \{(c')^3\}^2 - \{\mathbf{n}^3\}^2.$$

This is elementary: we have

$$e_3 = \langle e_3, c' \rangle c' + \langle e_3, \mathbf{n} \rangle \mathbf{n} + \langle e_3, \mathbf{b} \rangle \mathbf{b},$$

and when we take the inner product with e_3 we get

$$1 = \{(c')^3\}^2 + \{\mathbf{n}^3\}^2 + \{\mathbf{b}^3\}^2,$$

as desired.

Thus, when c is a curve with $\kappa = \tilde{\kappa}_g$ and $\tau^2 = -K$, we can find infinitely many isometries $\phi: \tilde{V} \rightarrow V$ with $\phi \circ \tilde{c} = c$; all the surfaces V are tangent to each other along c . In particular, if \tilde{c} is an asymptotic curve on M , with $K < 0$ along \tilde{c} , then we can take c to be \tilde{c} , and we see that a neighborhood of \tilde{c} can be continuously bent keeping \tilde{c} fixed; all surfaces in the bending are tangent to M along \tilde{c} . As opposed to this, if \tilde{c} satisfies $\tilde{\kappa}_g < \tilde{\kappa}$ everywhere, then there is only one other surface containing \tilde{c} which is isometric to a neighborhood of \tilde{c} in M , and it is nowhere tangent to M along \tilde{c} .

The case where $\kappa = \tilde{\kappa} = 0$ (both \tilde{c} and c are straight lines) is similar, except that now there is even complete leeway in the choice of the tangent space of V along c .

The discovery that asymptotic lines of a surface are precisely the curves along which the surface may be bent leads one to formulate all sorts of other questions.

For example, when is there an isometry $\phi: M \rightarrow \bar{M}$ which takes both families of asymptotic lines of M to asymptotic lines of \bar{M} ? It is easy to see that this happens essentially only when ϕ is the restriction of a Euclidean motion. In fact, if $f: U \rightarrow M$ is an imbedding for which the parameter lines are asymptotic curves, so that $l = n = 0$, and we define $\bar{f} = \phi \circ f$, then also $\bar{l} = \bar{n} = 0$. But we have, in addition,

$$ln - m^2 = \bar{l}\bar{n} - \bar{m}^2 \implies m = \pm\bar{m}.$$

If we restrict our attention to surfaces with $K < 0$, then we must have $m = \bar{m}$ or $m = -\bar{m}$ everywhere, and we can assume $m = \bar{m}$ by suitable choice of the normal. Hence ϕ is the restriction of a Euclidean motion.

Since this question turned out to be rather uninteresting, we modify it by investigating isometries $\phi: M \rightarrow \bar{M}$ which take the asymptotic curves of only one family of asymptotic lines on M to asymptotic lines on \bar{M} . Choose an orthonormal moving frame X_1, X_2 on M such that $\text{II}(X_1, X_1) = 0$, so that the integral curves of X_1 are the given family of asymptotic curves. Then we have

$$(1) \quad \begin{aligned} \psi_1^3 &= m\theta^2 \\ \psi_2^3 &= m\theta^1 + n\theta^2, \end{aligned}$$

where $0 = l = \text{II}(X_1, X_1)$ and $m = \text{II}(X_1, X_2)$ and $n = \text{II}(X_2, X_2)$. Let \bar{X}_1, \bar{X}_2 be the orthonormal moving frame $\bar{X}_i = \phi_* X_i$ on \bar{M} , and let barred forms (e.g., $\bar{\psi}_1^3$) actually denote ϕ^* of the corresponding forms on \bar{M} . Then $\bar{\theta}^i = \theta^i$ and $\bar{\omega}_2^1 = \omega_2^1$. Now $\bar{l} = \bar{\text{II}}(\bar{X}_1, \bar{X}_1) = 0$ by the hypothesis that our family of asymptotic curves is taken into asymptotic curves. Moreover,

$$ln - m^2 = \bar{l}\bar{n} - \bar{m}^2 \implies m = \pm\bar{m};$$

again we consider only the case $K < 0$ everywhere, so we might as well assume that $m = \bar{m}$, by suitable choice of the normals. Then we have

$$(2) \quad \begin{aligned} \bar{\psi}_1^3 &= m\theta^2 \\ \bar{\psi}_2^3 &= m\theta^1 + \bar{n}\theta^2. \end{aligned}$$

In particular,

$$\begin{aligned} \psi_1^3 = \bar{\psi}_1^3 &\implies d\psi_1^3 = d\bar{\psi}_1^3 \\ &\implies \omega_1^2 \wedge \psi_2^3 = \omega_1^2 \wedge \bar{\psi}_2^3 \quad \text{by Codazzi-Mainardi} \\ &\implies (n - \bar{n})\omega_1^2 \wedge \theta^2 = 0. \end{aligned}$$

Applying this to X_1, X_2 yields

$$(n - \bar{n}) \cdot \omega_1^2(X_1) = 0.$$

If $n = \bar{n}$ everywhere, then ϕ is a congruence. Assume that $n - \bar{n}$ is always $\neq 0$. Then $\omega_1^2(X_1) = 0 \implies \nabla_{X_1} X_1 = 0$, so the integral curves of X_1 are geodesics. Since they are also asymptotic curves, they must be straight lines; similarly their images, being both geodesics and asymptotic curves, are straight lines. In other words, this case involves a ruled surface being warped in such a way that the rulings remain straight.

[In general, it is easy to see that ruled surfaces can always be bent keeping their generators straight. In fact, suppose that our ruled surface is

$$f(s, t) = c(s) + t\delta(s), \quad |\delta| = 1 \implies \langle \delta, \delta' \rangle = 0.$$

Then

$$E = \langle c' + t\delta', c' + t\delta' \rangle = \langle c', c' \rangle + 2t\langle c', \delta' \rangle + t^2\langle \delta', \delta' \rangle$$

$$F = \langle c', \delta \rangle$$

$$G = 1.$$

Let $\bar{\delta}$ be any curve with

$$|\bar{\delta}| = 1, \quad |\bar{\delta}'| = |\delta'|.$$

In order for the surface

$$\bar{f}(s, t) = \bar{c}(s) + t\bar{\delta}(s)$$

to have the same metric as f , the curve \bar{c} must satisfy

$$|\bar{c}'| = |c'|$$

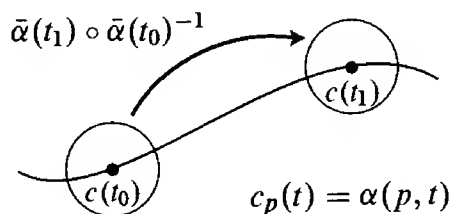
$$\langle \bar{c}', \bar{\delta}' \rangle = \langle c', \delta' \rangle$$

$$\langle \bar{c}', \bar{\delta} \rangle = \langle c, \delta \rangle,$$

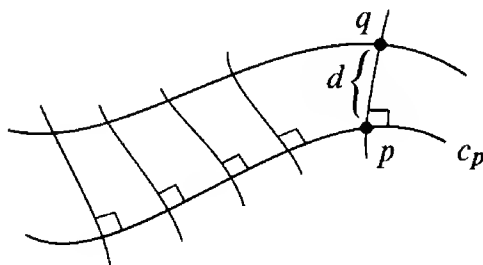
which is always solvable for \bar{c}' . The curve $\bar{\delta}$ can have essentially any shape. For example, if $\delta' \neq 0$ everywhere (the rulings are always changing), then we could reparameterize so that $|\delta| = |\delta'| = 1$. Then all we require is $|\bar{\delta}| = |\bar{\delta}'| = 1$, so that $\bar{\delta}$ can trace out any regular curve in S^2 .]

Here is one final classical problem about bendings. When does a surface M have a bending $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ such that each $\bar{\alpha}(t)(M) \subset M$? One example that immediately comes to mind is a surface of revolution. Obviously any

surface isometric to a surface of revolution also has this property. (The question really has almost nothing to do with surfaces in \mathbb{R}^3 , and is essentially intrinsic.) We will show that “in general” these are the only such surfaces. What we actually assume is that the various curves $c_p(t) = \alpha(t, p)$ give a foliation of M . Since each $\bar{\alpha}(t)$ is an isometry, each curve c_p has constant geodesic curvature, for $\bar{\alpha}(t_1) \circ \bar{\alpha}(t_0)^{-1}$ is an isometry taking a neighborhood of $c(t_0)$ to a neighborhood of $c(t_1)$. Now for any curve c_p , consider its “geodesic parallels”, the set of



points at a fixed distance d along the geodesics perpendicular to c_p . Let q be



the point on the geodesic intersecting c_p orthogonally at p . Clearly $\alpha(t, q)$ is on the geodesic intersecting c_p orthogonally at $\alpha(t, p)$. Thus the geodesic parallels of c_p are the other curves c_q . Note that the geodesics perpendicular to c_p are also perpendicular to c_q (Problem I.9-28). Now take a coordinate system u, v such that the v -parameter curves lie along the curves c_p , while the u -parameter curves are the arclength parameterized geodesics perpendicular to all curves c_p . Then the metric has the form

$$du \otimes du + G dv \otimes dv.$$

A computation shows that the geodesic curvature of the v -parameter curve through $(u, 0)$ is

$$\kappa_u(v) = -\frac{1}{2} \frac{G_u(u, v)}{G(u, v)} = -\frac{1}{2} \frac{\partial \log G}{\partial u}(u, v).$$

But $\kappa_u(v)$ depends only on u . So $\log G$ is of the form $a(u) + b(v)$, and thus G is of the form

$$G(u, v) = A(u) \cdot B(v).$$

Letting v_1 be a function with

$$v_1' = \sqrt{B},$$

our metric takes the form

$$du \otimes du + A(u) dv_1 \otimes dv_1.$$

Comparing with formula (4) on pg. III.158, we see that M is isometric to a surface of revolution. It doesn't seem worthwhile refining these purely local considerations by trying to analyze in detail just what happens when some of the curves c_p degenerate to points, but it certainly would be nice if one could prove that a compact surface M admitting a bending into itself is (globally) isometric to a surface of revolution.

Just for the hell of it, we will also look at a couple of classical local results about infinitesimal bendings. Let $M \subset \mathbb{R}^3$ be a surface, and let $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ be any variation of the inclusion map $i: M \rightarrow \mathbb{R}^3$ whose variation vector field Z at $t = 0$ is an infinitesimal bending of M . Let X_1, X_2 be an orthonormal moving frame on M . Then

$$\left. \frac{d}{dt} \right|_{t=0} \langle \bar{\alpha}(t)_* X_i, \bar{\alpha}(t)_* X_j \rangle = \langle dZ(X_i), X_j \rangle + \langle X_i, dZ(X_j) \rangle$$

by the argument on pages 171–172; since Z is an infinitesimal bending we thus have

$$(1) \quad \left. \frac{d}{dt} \right|_{t=0} \langle \bar{\alpha}(t)_* X_i, \bar{\alpha}(t)_* X_j \rangle = 0.$$

On each surface $\bar{\alpha}(t)(M)$ we can define an orthonormal moving frame $X_1(t), X_2(t)$ by applying the Gram-Schmidt orthonormalization process to $\bar{\alpha}(t)_* X_1, \bar{\alpha}(t)_* X_2$. Let $\theta^i(t)$ be $\bar{\alpha}(t)^*$ of the dual forms for this moving frame, so that $\theta^i(0) = \theta^i$, the dual forms for X_1, X_2 . Equation (1) is easily seen to imply that

$$(2) \quad 0 = \left. \frac{d}{dt} \right|_{t=0} \theta^i(t) = \dot{\theta}^i, \quad \text{say.}$$

For each t we have unique forms $\omega_j^i(t)$ with

$$(3) \quad \begin{aligned} \omega_j^i(t) &= -\omega_i^j(t) \\ d\theta^i(t) &= -\sum_{j=1}^2 \omega_j^i(t) \wedge \theta^j(t). \end{aligned}$$

Letting $t = 0$, we see that $\omega_j^i(t) = \omega_j^i$, the connection forms for X_1, X_2 . Now differentiate (3) with respect to t . Since (Problem 2) we always have

$$(d\eta)^\cdot = d\dot{\eta}$$

for any 1-parameter family of forms $\eta(t)$, we obtain, using (2),

$$\begin{aligned}\dot{\omega}_j^i &= -\dot{\omega}_i^j \\ 0 &= d\dot{\theta}^i = (d\theta^i)^\cdot = -\sum_{j=1}^2 \dot{\omega}_j^i \wedge \theta^j - \sum_{j=1}^2 \omega_j^i \wedge \dot{\theta}^j \\ &= -\sum_{j=1}^2 \dot{\omega}_j^i \wedge \theta^j.\end{aligned}$$

It follows that $\dot{\omega}_2^1 = 0$. Now we differentiate the equation

$$d\omega_2^1(t) = K(t) \theta^1(t) \wedge \theta^2(t),$$

to obtain

$$0 = d\dot{\omega}_2^1 = (d\omega_2^1)^\cdot = \dot{K} \theta^1 \wedge \theta^2 + 0.$$

Thus we see that we always have

$$\dot{K} = 0.$$

It now seems a natural enough question to ask when we have $\dot{H} = 0$. The answer to this question is left to Problem 4.

Another question, especially interesting in view of Lemma 6, is to find those surfaces M which possess an infinitesimal bending Z that is everywhere *tangent* to M . Once again, surfaces of rotation are obvious examples. Moreover, one can easily show that if Z is an infinitesimal bending of M which is tangent to M , and $f: M \rightarrow \bar{M}$ is an isometry, then $\bar{Z} = f_*Z$ is an infinitesimal bending of \bar{M} .

Again we can show that “in general” these are the only such surfaces. Given a nowhere 0 infinitesimal bending Z tangent along M , we choose an immersion $f: U \rightarrow M$ such that the v -parameter curves lie along the integral curves of Z , and the u -parameter curves along the curves perpendicular to them. Then $I_f = f^*(\ , \)$ has the form

$$I_f = E du \otimes du + G dv \otimes dv.$$

By assumption, Z is always proportional to $\partial f / \partial v$, and it will be convenient to write Z as

$$Z = \frac{\lambda}{\sqrt{G}} \frac{\partial f}{\partial v} = \frac{\lambda}{\sqrt{G}} f_2.$$

Then the equations

$$0 = \langle Z_1, f_1 \rangle = \langle Z_2, f_2 \rangle, \quad 0 = \langle Z_1, f_2 \rangle + \langle Z_2, f_1 \rangle$$

lead to

$$\frac{\partial E}{\partial v} = 0, \quad \frac{\partial \lambda}{\partial v} = 0, \quad \frac{\partial \left(\frac{\lambda}{\sqrt{G}} \right)}{\partial v} = 0.$$

From the first we see that we can alter the immersion f so that $E = 1$ everywhere. From the third we see that we can likewise arrange that $\lambda = \sqrt{G}$ everywhere. Then our metric has the form

$$du \otimes du + \lambda(u)^2 dv \otimes dv,$$

as desired.

Most of these local results are rather unsatisfying, since they usually require some subsidiary conditions of the same nature as those used in the classical classification of flat surfaces. But there are certain questions where local results are precisely what we should be interested in. We have already seen (pages 209–210) that there are isometric compact surfaces in \mathbb{R}^3 which cannot be connected by a bending. But it seems likely that isometric surfaces can *locally* be connected by a bending. Actually, the argument on page 210 shows that even this is false, since a surface of positive curvature can never be bent into its mirror image. So we should instead conjecture that given any two isometric surfaces, the first can locally be connected by a bending to the second or else to its mirror image.

To investigate this question, we consider once again the Darboux equation. From the considerations on pages 143–146 we see that the immersions $f = (u, v, w): U \rightarrow \mathbb{R}^3$ defined in a neighborhood of $0 \in \mathbb{R}^2$ such that

- (i) $I_f = E dx \otimes dx + F[dx \otimes dy + dy \otimes dx] + G dy \otimes dy$
- (ii) $w(0) = w_1(0) = w_2(0) = 0$
- (iii) $u(0) = v(0) = 0$
- (iv) $u_1(0) = 0, \quad u_2(0) > 0, \quad v_1(0) > 0$

are in one-one correspondence with the solutions w of the Darboux equation which satisfy (ii). Writing the Darboux equation as on page 146 [equation (6)], we see that the following holds:

- (*) Let E, F, G be the components of a metric in a neighborhood of $0 \in \mathbb{R}^2$, and let ρ, σ be two functions in a neighborhood of $0 \in \mathbb{R}$ with

$$\begin{aligned}\rho(0) &= \rho'(0) = \sigma(0) = 0 \\ \rho''(0) &\neq 0.\end{aligned}$$

Assume E, F, G and ρ, σ are analytic, unless the curvature K satisfies $K(0) < 0$. Then there is a unique immersion $f = (u, v, w)$ defined in a neighborhood of $0 \in \mathbb{R}^2$ such that (i)–(iv) hold, and for which

$$w(x, 0) = \rho(x), \quad w_2(x, 0) = \sigma(x).$$

From this observation it is but a short step to

18. LEMMA. Let $\phi: M \rightarrow \bar{M}$ be an isometry between two surfaces $M, \bar{M} \subset \mathbb{R}^3$ and let $X \in M_p$ be a vector such that $\text{II}(X_p, X_p)$ and $\bar{\text{II}}(\phi_* X_p, \phi_* X_p)$ are either both positive or both negative. Assume that M and \bar{M} are analytic surfaces, unless $K(p) < 0$. Then there is a neighborhood U of p and a bending $\alpha: [0, 1] \times U \rightarrow \mathbb{R}^3$ with $\bar{\alpha}(0) = \text{identity}$ and $\bar{\alpha}(1) = \phi$.

PROOF. Choose immersions $f = (u, v, w)$ and $\bar{f} = (\bar{u}, \bar{v}, \bar{w})$ taking a neighborhood of $0 \in \mathbb{R}^2$ into M and \bar{M} , respectively, with $f(0) = p$ and $\bar{f}(0) = \phi(p)$. Without loss of generality we can assume that both f and \bar{f} satisfy the conditions (ii)–(iv) above, and that $X_p = f_*((1, 0))$ and $\phi_* X_p = \bar{f}_*((1, 0))$. Let $I_f = I_{\bar{f}}$ have components E, F, G . For $0 \leq t \leq 1$, let

$$\begin{aligned}\rho_t(x) &= (1 - t)w(x, 0) + t\bar{w}(x, 0) \\ \sigma_t(x) &= (1 - t)w_2(x, 0) + t\bar{w}_2(x, 0).\end{aligned}$$

Then (ii) gives

$$(1) \quad \rho_t(0) = \rho'_t(0) = \sigma_t(0) = 0.$$

If l, m, n and $\bar{l}, \bar{m}, \bar{n}$ are the coefficients of II_f and $\text{II}_{\bar{f}}$, then by assumption $\text{II}(X_p, X_p) = l(0, 0)$ and $\bar{\text{II}}(\phi_* X_p, \phi_* X_p) = \bar{l}(0, 0)$ have the same sign. Now

$$\begin{aligned}l(0, 0) &= \frac{1}{\sqrt{EG - F^2}} \cdot \det \begin{pmatrix} f_{11} \\ f_1 \\ f_2 \end{pmatrix} \quad \text{at } (0, 0) \quad [\text{formula (A) of Chapter 3}] \\ &= \frac{1}{\sqrt{EG - F^2}} \cdot \det \begin{pmatrix} u_{11} & v_{11} & w_{11} \\ u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \end{pmatrix} \quad \text{at } (0, 0)\end{aligned}$$

$$= \frac{1}{\sqrt{EG - F^2}}(0, 0) \cdot w_{11}(0, 0) \cdot u_2(0, 0) \cdot v_1(0, 0) \quad \text{by (ii) and (iv),}$$

and similarly for $\bar{l}(0, 0)$. Using (iv), we see that $w_{11}(0)$ has the same sign as $l(0, 0)$, and similarly for $\bar{w}_{11}(0)$; so $w_{11}(0)$ and $\bar{w}_{11}(0)$ have the same sign. It follows that

$$(2) \quad \rho_t''(0) \neq 0.$$

By (*), there are unique immersions $f_t = (u_t, v_t, w_t)$ defined in a neighborhood of $0 \in \mathbb{R}^2$ such that (i)–(iv) hold, and for which

$$w_t(x, 0) = \rho_t(x), \quad (w_t)_2(x, 0) = \sigma_t(x).$$

The uniqueness implies that $f_0 = f$ and $f_1 = \bar{f}$.

The only details which need to be checked are that all f_t can be defined in a common neighborhood of $0 \in \mathbb{R}^2$, and that the f_t vary smoothly with t . This unrewarding task is left to the reader. ♦

19. THEOREM (E. E. LEVI). Let $\phi: M \rightarrow \bar{M}$ be an isometry between two surfaces $M, \bar{M} \subset \mathbb{R}^3$, and suppose that $p \in M$ and $\phi(p) \in \bar{M}$ are not planar points. Assume that M and \bar{M} are analytic surfaces, unless $K(p) < 0$. Then there is a neighborhood U of p and a bending $\alpha: [0, 1] \times U \rightarrow \mathbb{R}^3$ with $\bar{\alpha}(0) = \text{identity}$ and either $\bar{\alpha}(1) = \phi$ or $\bar{\alpha}(1) = R \circ \phi$, where R is a reflection.

PROOF. Since p and $\phi(p)$ are not planar points, there are at most 2 asymptotic directions at these points, and thus certainly a tangent vector $X \in M_p$ such that $\text{II}(X, X)$ and $\bar{\text{II}}(\phi_*X, \phi_*X)$ are both non-zero. If they have the same sign we apply Lemma 18; if they have opposite signs we apply Lemma 18 to M and $R(\bar{M})$. ♦

We know that the reflection R has to be allowed if $K(p) > 0$. It turns out that R is unnecessary if $K(p) < 0$. We begin with a preliminary result.

20. LEMMA. Let E, F, G be the components of a metric in a neighborhood of $0 \in \mathbb{R}^2$, and $\mathbf{v} \in \mathbb{R}^2_0$ a given tangent vector. Assume that E, F, G are analytic, unless $K < 0$. Then there is an immersion $f = (u, v, w)$ in a neighborhood of 0 such that I_f has components E, F, G , and $f_*(\mathbf{v})$ is not an asymptotic vector on the image of f .

PROOF. Without loss of generality, we can assume that $\mathbf{v} = (1, 0)$. Then choose any two functions ρ, σ satisfying

$$\begin{aligned}\rho(0) &= \rho'(0) = \sigma(0) = 0 \\ \rho''(0) &\neq 0,\end{aligned}$$

and consider the immersion f determined by $(*)$, with

$$\begin{aligned}u_2(0) &> 0, & v_1(0) &> 0 \\ w(x, 0) = \rho(x) &\implies w_{11}(0, 0) = \rho''(0).\end{aligned}$$

The calculation in the proof of Lemma 18 shows that

$$l(0, 0) = \frac{1}{\sqrt{EG - F^2}}(0, 0) \cdot w_{11}(0, 0) \cdot u_2(0, 0) \cdot v_1(0, 0) \neq 0.$$

This means that $f_*(\mathbf{v})$ is not an asymptotic vector. \blacklozenge

21. THEOREM (E. E. LEVI). Let $\phi: M \rightarrow \bar{M}$ be an isometry between two surfaces $M, \bar{M} \subset \mathbb{R}^3$, and suppose that $K(p) < 0$. Then there is a neighborhood U of p and a bending $\alpha: [0, 1] \times U \rightarrow \mathbb{R}^3$ with $\bar{\alpha}(0) = \text{identity}$ and $\bar{\alpha}(1) = \phi$.

PROOF. We just have to show that a neighborhood of p in M can be bent into its mirror image $R(M)$. Let $X \in M_p$ be an asymptotic vector. Then there are vectors Y arbitrarily close to X in M_p with $\text{II}(Y, Y) > 0$, as well as vectors arbitrarily close to X with $\text{II}(Y, Y) < 0$. Lemma 20 says that there is (locally) an isometry $\psi: M \rightarrow \tilde{M} \subset \mathbb{R}^3$ such that $\psi_*(X)$ is *not* an asymptotic vector on \tilde{M} . We can assume, by composing \tilde{M} with a reflection if necessary, that $\tilde{\text{II}}(\psi_*X, \psi_*X) > 0$. Then the same inequality holds for all tangent vectors of $\tilde{M}_{\psi(p)}$ in some sector containing ψ_*X . So we can choose $Y \in M_p$ with

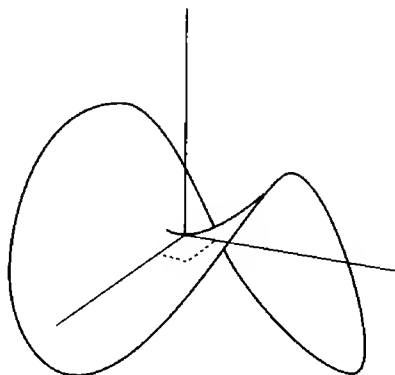
$$\text{II}(Y, Y) > 0, \quad \tilde{\text{II}}(\psi_*Y, \psi_*Y) > 0,$$

and it follows from Theorem 19 that there is a bending of a neighborhood of p in M onto a neighborhood of $\psi(p)$ in \tilde{M} . But we can also choose $Y \in M_p$ with

$$\text{II}(Y, Y) < 0, \quad \tilde{\text{II}}(\psi_*Y, \psi_*Y) > 0,$$

and then it follows that there is a bending of a neighborhood of $R(p)$ in $R(M)$ onto a neighborhood of $\psi(p)$ in \tilde{M} . Consequently, there is a bending of a neighborhood of p in M onto a neighborhood of $R(p)$ in $R(M)$. \blacklozenge

This little proof, clever as it is, certainly doesn't give any idea of what is going on geometrically. E. E. Levi supplied a geometric description of the bending in the special case of a surface M of constant negative curvature whose asymptotic directions are perpendicular at $p \in M$. Rotation through an angle of $\pi/2$

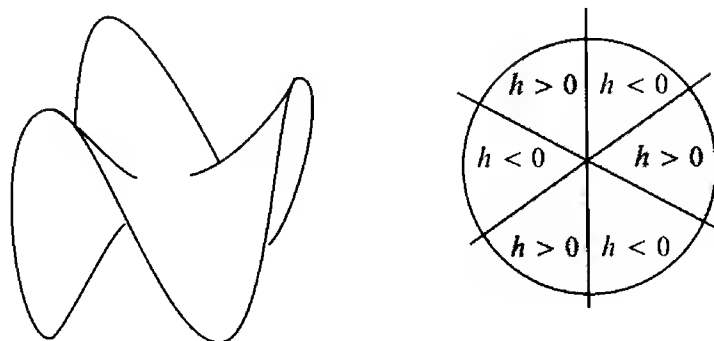


around the normal at p takes M into its reflection R through the tangent plane at p . But the isometry $M \rightarrow R(M)$ thus obtained is not the same as $R|M$. To modify this, we consider the series of isometries obtained as follows. At time t , we first perform a rotation A_t through an angle t around the normal, and then compose $A_t|M$ with a map $B_t: A_t(M) \rightarrow A_t(M)$ of the surface of constant curvature $A_t(M)$ onto itself which rotates the tangent space M_p back by an angle of $-t$. For $t = \pi/2$ we obtain the map $R|M$.

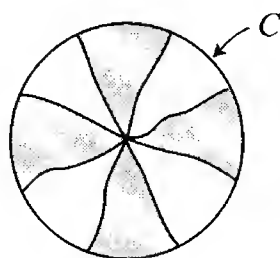
These results of E. E. Levi [1] were proved nearly 30 years after A. Voss had first explicitly pointed out that a distinction ought to be made between warpings and bendings. Levi's results were regarded as demonstrations that these distinctions really did not exist (at least locally). Of course, Theorem 19 does have the added hypothesis that p and $\phi(p)$ are not planar points; in E. E. Levi's original theorem, there was the stronger requirement that $K(p) \neq 0$. Such requirements were regarded, if they were regarded at all, as merely technical details. Remarkably enough, H. Schilt [1] discovered that Theorem 19 is actually false if the point p is a flat point, even if all points in a neighborhood of p have $K < 0$. We will outline the arguments here, but for some of the details the reader is referred to Schilt's paper, which is very clearly written and easy to follow.

Consider a surface M which is the graph of a function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ with $0 = h(0) = h_1(0) = h_2(0)$. If the curvature $K(0) < 0$, then M looks like a "saddle", as we saw in Chapter 2. But if $K(0) = 0$, and $K < 0$ in a deleted neighborhood of 0, then it can be shown that M looks like a "generalized

monkey saddle”: inside a sufficiently small circle C around 0, the zero set of h

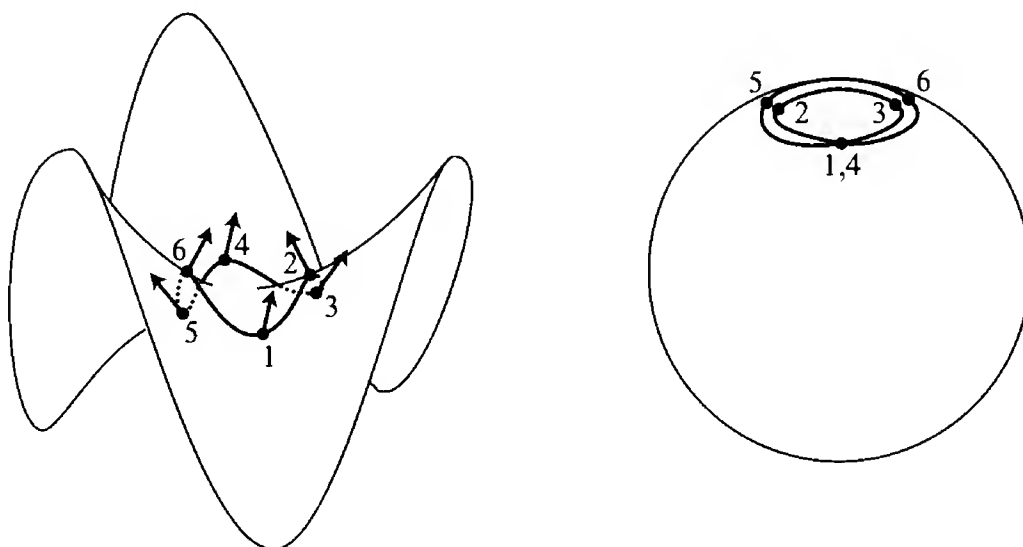


consists of an even number 2λ of curves starting at 0 and ending at C , with no points in common except 0; the sign of h on the sectors between these curves

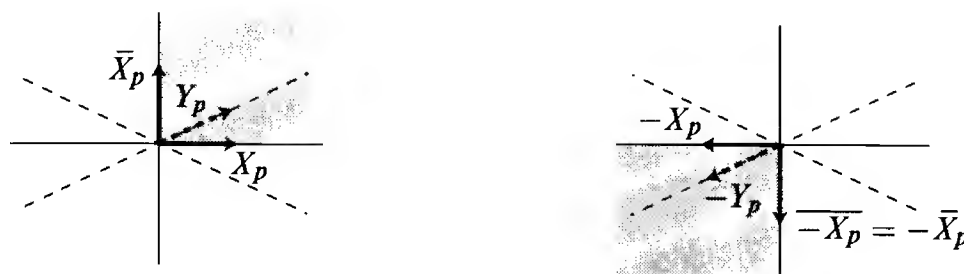


is constant and changes as we go from one sector to another. The number $s = \lambda - 1$ is called the **order** of the saddle point at 0. An ordinary saddle point, where $K(0) < 0$, has order 1, while the monkey saddle has order 2. The graph of $h(x, y) = \operatorname{Re}(x + iy)^{s+1}$ has order s .

The order of a saddle point can be described in another way, by considering a closed curve in M going once around p in the positively oriented sense. It turns out that the image of this curve under the normal map goes s times around the normal at p , but in the negatively oriented sense. The following picture illustrates this for the monkey saddle.

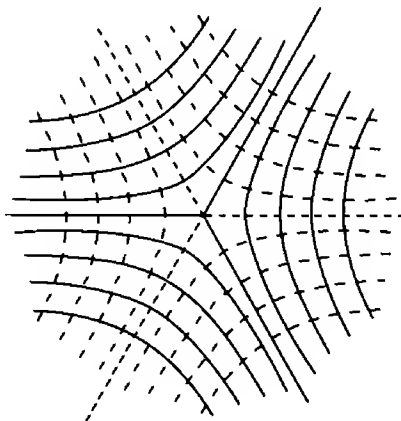


Finally, the order of a saddle point may be described in yet a third way. In the region where $K < 0$, the principal curvatures k_1, k_2 are of different signs, say $k_1 > 0 > k_2$. So we can pick out the 1-dimensional distribution of all multiples of the principal directions corresponding to the principal curvature k_1 , say. Notice that if $X_p \in M_p$ is a principal direction with principal curvature k_1 , and \bar{X}_p is a perpendicular vector with (X_p, \bar{X}_p) positively oriented, then there is just one unit asymptotic vector Y_p in the quadrant of M_p bounded by X_p and \bar{X}_p . Moreover, if we start with $-X_p$ instead of X_p , then we just end up



with $-Y_p$. Thus we can also pick out a 1-dimensional distribution consisting of all multiples of an asymptotic vector at each point.

Another way of stating these facts is the following: the principal curves and the asymptotic curves can each be separated into two distinct families in the region where $K < 0$.^{*} The following picture shows the projections on the (x, y) -plane of the two families of asymptotic lines (one indicated by solid lines, the other by dotted lines) for the ordinary monkey saddle.



^{*}Schilt [1] tries to do this for the asymptotic curves by looking at the signs of their torsions, which have to be different by the Beltrami-Enneper theorem. But this doesn't work at points where the torsions don't exist.

Now one can define the index of any one of these distributions (Addendum 2 to Chapter 4); this index i is a half-integer, and is clearly the same for both families. It is related to the order s of the saddle by

$$(*) \quad s = 1 - 2i.$$

For example, the monkey saddle, with $s = 2$, has $i = -1/2$ (compare the above picture with the one on pg. III.219). On the other hand, if 0 is a parabolic point, then the asymptotic curves have no singularity at 0, so $i = 0$ and $s = 1$, just as in the case of an ordinary saddle point.

Proving $(*)$ is not completely straightforward, but the relation is very important, for it leads immediately to the result that s is a bending invariant: given a bending $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ of the inclusion map of M into \mathbb{R}^3 , each surface $\bar{\alpha}(t)(M)$ has a saddle at $\alpha(t, 0)$, and for all t the order of this saddle is s . For the proof one merely observes that the principal curves or asymptotic curves vary continuously, and therefore always have the same index.

One might think that this result is hardly worth mentioning, on the grounds that s should actually be a warping invariant: any surface isometric to M also ought to have a saddle of order s . But this is *not* true! For suppose that M is an analytic surface with a saddle of order $s > 1$. By Lemma 20, there is a surface \bar{M} isometric to M such that the point corresponding to p is a parabolic point. Thus the saddle order at this point is $\bar{s} = 1$. Consequently, there is no bending from M to \bar{M} ; indeed, no neighborhood of the saddle point on M can be bent onto its isometric image in \bar{M} !

These examples of isometric surfaces which are not even locally connected by a bending all have $K < 0$ in a neighborhood of the point with $K = 0$. But Hopf and Schilt [1] show that, for certain classes of surfaces, the order of contact of the graph of h with the (x, y) -plane is also a bending invariant, but not a warping invariant. This allows them to give examples of the same phenomenon, but for surfaces with $K = 0$ at one point and $K > 0$ in a neighborhood of the point. They are also able to show that Theorem 21 fails if $K(p) = 0$, even if $K < 0$ for all other points in a neighborhood of p .

It should be mentioned that the present proofs of Theorems 19 and 21 (which slightly strengthen E. E. Levi's original result) are due to Schilt, who also observed that the above analysis of the Darboux equation can be used to prove the following: If $p \in M$ is not a planar point, then [assuming M is analytic, unless $K(p) < 0$] some neighborhood of p has a non-trivial bending. Actually, this result holds without assumptions of analyticity when $K(p) > 0$, although the proof is much harder; in fact, this comes out of the proof that a convex surface with a disc deleted is bendable (see Hellwig [1]). However, the case $K(p) = 0$

is unresolved. There is a startling result of Efimov along these lines (see Efimov {1}, Chapter IX, and Hoesli [1]): There exist infinitely many examples of analytic surfaces containing a point p such that no neighborhood of p has any non-trivial analytic bendings; in fact, any smooth bending of this neighborhood into analytic surfaces is trivial. A specific example is $\{(x, y, z) : z = (x^2 + y^2)^5\}$ with $p = 0 \in \mathbb{R}^3$. Whether there are examples where no neighborhood has any non-trivial bendings into C^∞ surfaces is unknown, and certainly a highly intriguing question. As far as I can tell, it is not even known whether every C^∞ surface is locally warpable.

With these considerations we finally end our investigation of rigidity for surfaces in \mathbb{R}^3 . So far we have said absolutely nothing about surfaces in S^3 or H^3 . Recollection of Chapter 7F might make even the stoutest hearts quail at this prospect, but fortunately there is an incredibly neat trick, due to Pogorelov {3}, which reduces almost all such questions to the case of surfaces in \mathbb{R}^3 . Since Pogorelov's book is written more in the style of the Russian school, and includes many results specifically tailored for such a study, we will give a treatment of the main points totally from the C^∞ point of view.

In Chapter 7A we considered the central projection $\phi: S^{n+} \rightarrow \mathbb{R}^n$, from the open northern hemisphere S^{n+} of S^n onto \mathbb{R}^n . It is easily computed that

$$\phi(x) = \left(\frac{x_1}{x_{n+1}}, \dots, \frac{x_n}{x_{n+1}} \right) \quad \text{for } x \in S^{n+}.$$

Let $*$ = $(0, \dots, 0, 1)$ be the north pole of S^n . Then ϕ can also be described by

$$\phi(x) = \frac{x - \langle x, * \rangle \cdot *}{\langle x, * \rangle} \in \mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+1}.$$

Now let $f_1, f_2: M \rightarrow S^n$ be two maps of a Riemannian manifold M into S^n . Define $\bar{f}_1: M \rightarrow \mathbb{R}^n$ (actually, into $\mathbb{R}^n \times \{0\}$) by

$$(*) \quad \bar{f}_1(p) = \frac{f_1(p) - \langle f_1(p), * \rangle \cdot *}{\langle f_1(p) + f_2(p), * \rangle},$$

and define \bar{f}_2 similarly (this formula makes sense so long as $f_1 + f_2$ is never perpendicular to $*$, which happens, in particular, if f_1 and f_2 both go into S^{n+}).

22. PROPOSITION. The two maps $f_1, f_2: M \rightarrow S^n$ induce the same (possibly degenerate) metric on M if and only if the two maps $\bar{f}_1, \bar{f}_2: M \rightarrow \mathbb{R}^n$ induce the same metric on M .

PROOF. For the \mathbb{R}^n -valued form df_1 we compute from equation (*) that

$$(1) \quad \langle f_1 + f_2, * \rangle^2 df_1 = \langle f_1 + f_2, * \rangle [df_1 - \langle df_1, * \rangle \cdot *] \\ - \langle df_1 + df_2, * \rangle \cdot [f_1 - \langle f_1, * \rangle \cdot *].$$

[This equation means that

$$(2) \quad \langle f_1(p) + f_2(p), * \rangle^2 df_1(X) \\ = \langle f_1(p) + f_2(p), * \rangle [df_1(X) - \langle df_1(X), * \rangle \cdot *] \\ - \langle df_1(X) + df_2(X), * \rangle \cdot [f_1(p) - \langle f_1(p), * \rangle \cdot *]$$

for all $X \in M_p$.]

Since $\langle f_1, f_1 \rangle = 1 \implies \langle df_1, f_1 \rangle = 0$, we have

$$\begin{aligned} \langle f_1 + f_2, * \rangle^4 |df_1|^2 &= \langle f_1 + f_2, * \rangle^2 [|df_1|^2 - \langle df_1, * \rangle^2] \\ &\quad + 2\langle f_1 + f_2, * \rangle \langle df_1 + df_2, * \rangle \langle f_1, * \rangle \langle df_1, * \rangle \\ &\quad + \langle df_1 + df_2, * \rangle^2 [1 - \langle f_1, * \rangle^2] \\ &= \langle f_1 + f_2, * \rangle^2 |df_1|^2 + \langle df_1 + df_2, * \rangle^2 \\ &\quad - [\langle f_1 + f_2, * \rangle \langle df_1, * \rangle - \langle df_1 + df_2, * \rangle \langle f_1, * \rangle]^2 \\ &= \langle f_1 + f_2, * \rangle^2 |df_1|^2 + \langle df_1 + df_2, * \rangle^2 \\ &\quad - [\langle f_2, * \rangle \langle df_1, * \rangle - \langle df_2, * \rangle \langle f_1, * \rangle]^2, \end{aligned}$$

which is symmetric in f_1 and f_2 if and only if $|df_1|^2 = |df_2|^2$. ♦

Remark: Even if f_1 and f_2 are immersions, the maps \bar{f}_1 and \bar{f}_2 need not be. However, if $*$ is not a linear combination of $f_1(p)$ and any vector $df_1(X)$ for $X \in M_p$, then the vectors $df_1(x) - \langle df_1(X), * \rangle *$ and $f_1(p) - \langle f_1(p), * \rangle *$ in (2) are linearly independent, so $d\bar{f}_1(X) \neq 0$ for all $X \in M_p$, and \bar{f}_1 is an immersion at p .

Suppose we have two maps $f_1, f_2: M \rightarrow \mathbb{R}^n \subset \mathbb{R}^{n+1}$. We define $\hat{f}_1: M \rightarrow S^n$ by

$$\hat{f}_1(p) = \frac{2f_1(p) + (1 - |f_1(p)|^2 + |f_2(p)|^2) \cdot *}{|2f_1(p) + (1 - |f_1(p)|^2 + |f_2(p)|^2) \cdot *|},$$

and we define $\hat{f}_2: M \rightarrow S^n$ similarly. It is easy to compute that if we begin with two maps $f_1, f_2: M \rightarrow S^n$, form $\bar{f}_1, \bar{f}_2: M \rightarrow \mathbb{R}^n$ as before, and then apply the present construction to \bar{f}_1, \bar{f}_2 , obtaining $\hat{\bar{f}}_1, \hat{\bar{f}}_2: M \rightarrow S^n$, then

$$\hat{\bar{f}}_1 = f_1, \quad \hat{\bar{f}}_2 = f_2.$$

Similarly, if we start with $f_1, f_2: M \rightarrow \mathbb{R}^n$, then

$$\bar{\bar{f}}_1 = f_1, \quad \bar{\bar{f}}_2 = f_2.$$

Hence

23. COROLLARY. The two maps $f_1, f_2: M \rightarrow \mathbb{R}^n$ induce the same metric on M if and only if the two maps $\hat{f}_1, \hat{f}_2: M \rightarrow S^n$ induce the same metric on M .

On the other hand, our new constructions also preserve the notion of congruence.

24. PROPOSITION. Let A be an orthogonal map of \mathbb{R}^{n+1} , so that $A: S^n \rightarrow S^n$ is an isometry, and let $f_1, f_2: M \rightarrow S^n$ be maps with $f_2 = A \circ f_1$. Suppose that we can define $\bar{f}_1, \bar{f}_2: M \rightarrow \mathbb{R}^n$. Then there is a Euclidean motion $\bar{A}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\bar{f}_2 = \bar{A} \circ \bar{f}_1$.

PROOF. We just have to show that

$$|\bar{f}_1(p) - \bar{f}_1(q)|^2 = |\bar{f}_2(p) - \bar{f}_2(q)|^2$$

for all $p, q \in M$. For any x with $x + Ax$ not perpendicular to $*$, let $x^\# \in \mathbb{R}^{n+1}$ be a multiple of x such that

$$(I) \quad \langle x^\# + Ax^\#, * \rangle = 1.$$

Then

$$\bar{f}_1(p) = \frac{f_1(p) - \langle f_1(p), * \rangle *}{\langle f_1(p) + A(f_1(p)), * \rangle} = f_1(p)^\# - \langle f_1(p)^\#, * \rangle \cdot *,$$

and similarly for $\bar{f}_1(q), \bar{f}_2(p), \bar{f}_2(q)$. Set $z = f_1(p)^\# - f_1(q)^\#$. Then

$$\begin{aligned} |\bar{f}_1(p) - \bar{f}_1(q)|^2 &= |z - \langle z, * \rangle \cdot *|^2 = |z|^2 - \langle z, * \rangle^2 \\ |\bar{f}_2(p) - \bar{f}_2(q)|^2 &= |Az - \langle Az, * \rangle \cdot *|^2 = |Az|^2 - \langle Az, * \rangle^2. \end{aligned}$$

But $|Az| = |z|$, while

$$\begin{aligned} \langle z + Az, * \rangle &= \langle f_1(p)^\# + Af_1(p)^\# - [f_1(q)^\# + Af_1(q)^\#], * \rangle = 0, \quad \text{by (I)} \\ \implies \langle z, * \rangle^2 &= \langle Az, * \rangle^2. \quad \spadesuit \end{aligned}$$

25. PROPOSITION. Let $B: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an isometry, and let $f_1, f_2: M \rightarrow \mathbb{R}^n$ be maps with $f_2 = B \circ f_1$. Then there is an isometry $A: S^n \rightarrow S^n$ such that $\hat{f}_2 = A \circ \hat{f}_1$.

PROOF. Define $\rho_1, \rho_2: \mathbb{R}^n \rightarrow S^n$ by

$$\rho_1(y) = \frac{2y + (1 - |y|^2 + |By|^2) \cdot *}{|\text{numerator}|}$$

$$\rho_2(y) = \frac{2y + (1 - |y|^2 + |B^{-1}y|^2) \cdot *}{|\text{numerator}|},$$

so that

$$\hat{f}_1 = \rho_1 \circ f_1, \quad \hat{f}_2 = \rho_2 \circ f_2.$$

It is clear that ρ_1 and ρ_2 are continuous. We claim that ρ_1 and ρ_2 are one-one. Suppose instead that $y \neq z \in \mathbb{R}^n$, but $\rho_1(y) = \rho_1(z)$. Clearly y and z must be linearly dependent, so there is a unit vector $v \in \mathbb{R}^n$ with $y = \lambda v$ and $z = \mu v$. Then $\rho_1(y) = \rho_1(z)$ implies that

$$\frac{2\lambda}{|\text{numerator for } \rho_1(y)|} = \frac{2\mu}{|\text{numerator for } \rho_1(z)|}$$

$$\frac{1 - \lambda^2 + |By|^2}{|\text{numerator for } \rho_1(y)|} = \frac{1 - \mu^2 + |Bz|^2}{|\text{numerator for } \rho_1(z)|},$$

and hence

$$(1) \quad \frac{1 - \lambda^2 + |By|^2}{\lambda} = \frac{1 - \mu^2 + |Bz|^2}{\mu}.$$

Let $B = T_w \circ C$, where C is a rotation, and T_w is translation by a vector w . Then

$$|By|^2 = |C(\lambda v) + w|^2 = \lambda^2 + 2\langle C(\lambda v), w \rangle + |w|^2$$

$$|Bz|^2 = \mu^2 + 2\langle C(\mu v), w \rangle + |w|^2.$$

So (1) becomes

$$\frac{1 + |w|^2}{\lambda} + 2\langle C(v), w \rangle = \frac{1 + |w|^2}{\mu} + 2\langle C(v), w \rangle.$$

Hence $\lambda = \mu \implies y = z$. Similarly, ρ_2 is one-one.

Since ρ_1 and ρ_2 are continuous one-one maps between manifolds of the same dimension, their images are open, by Invariance of Domain (Theorem I.1-1). So we can consider the map $\rho_2 \circ B \circ \rho_1^{-1}$, defined on some open set in S^n . We claim that this map preserves distances on S^n , and is thus the restriction of some isometry A . Since

$$\hat{f}_2 = \rho_2 \circ f_2 = \rho_2 \circ B \circ f_1 = \rho_2 \circ B \circ \rho_1^{-1} \circ \rho_1 \circ f_1 = \rho_2 \circ B \circ \rho_1^{-1} \circ \hat{f}_1,$$

this will prove the Theorem.

It suffices to show that

$$|\rho_1(y) - \rho_1(z)|^2 = |\rho_2(By) - \rho_2(Bz)|^2,$$

since the distance between two points in S^n is determined by their Euclidean distance. Clearly, we just have to show that for all y and z we have

$$\langle \rho_1(y), \rho_1(z) \rangle = \langle \rho_2(By), \rho_2(Bz) \rangle.$$

If

$$\begin{aligned} a(y) &= 2y + (1 - |y|^2 + |By|^2) \cdot * \\ b(y) &= 2By + (1 - |By|^2 + |y|^2) \cdot *, \end{aligned}$$

then

$$\rho_1(y) = \frac{a(y)}{|a(y)|}, \quad \rho_2(By) = \frac{b(y)}{|b(y)|}.$$

So it suffices to show that for all $y, z \in \mathbb{R}^n$ we have

$$(2) \quad \langle a(y), a(z) \rangle = \langle b(y), b(z) \rangle.$$

Now

$$(3) \quad \begin{cases} \langle a(y), a(z) \rangle = 4\langle y, z \rangle + (1 - |y|^2 + |By|^2)(1 - |z|^2 + |Bz|^2) \\ \langle b(y), b(z) \rangle = 4\langle By, Bz \rangle + (1 - |By|^2 + |y|^2)(1 - |Bz|^2 + |z|^2). \end{cases}$$

Writing $B = T_w \circ C$ as before, we have

$$\begin{aligned} |By|^2 &= |y|^2 + 2\langle Cy, w \rangle + |w|^2 \\ |Bz|^2 &= |z|^2 + 2\langle Cz, w \rangle + |w|^2, \end{aligned}$$

and

$$\langle By, Bz \rangle = \langle y, z \rangle + \langle Cy + Cz, w \rangle + |w|^2.$$

Substituting into (3), we obtain (2). ♦

The hardest problem is to show that our construction preserves convexity. This holds only in certain circumstances, which requires a preliminary remark. We say that a hypersurface $M \subset S^{n+}$ is **star-shaped with respect to $*$** if each geodesic ray starting from $*$, and contained in S^{n+} , intersects M exactly once. Clearly M has a natural orientation, just as in the case of hypersurfaces $M \subset \mathbb{R}^n$ that are star-shaped with respect to 0. Our results will hold only for imbeddings $f_1, f_2: M \rightarrow S^{n+}$ or \mathbb{R}^n whose images are star-shaped with respect to $*$ or 0; moreover, for some orientation on M , the induced orientations on $f_1(M)$ and $f_2(M)$ must be the natural ones.

Before giving the precise results, we consider one more preliminary. Let γ be an arclength parameterized curve in S^n . Then the Frenet equations for S^n give

$$\begin{aligned}\kappa(s)\mathbf{n}(s) &= \frac{D\gamma'(s)}{ds} = \mathbf{T}\gamma''(s) \\ &= \gamma''(s) - \langle \gamma''(s), \gamma(s) \rangle \cdot \gamma(s).\end{aligned}$$

But

$$\langle \gamma, \gamma \rangle = 1 \implies \langle \gamma', \gamma \rangle = 0 \implies \langle \gamma'', \gamma \rangle = -\langle \gamma', \gamma' \rangle = -1.$$

So we have

$$\gamma''(s) = \kappa(s)\mathbf{n}(s) - \gamma(s).$$

Hence

$$\begin{aligned}(\ast) \quad \gamma(s+h) &= \gamma(s) + h\gamma'(s) + \frac{h^2}{2}\gamma''(s) + o(h^2) \\ &= \left(1 - \frac{h^2}{2}\right)\gamma(s) + h\mathbf{t}(s) + \frac{h^2}{2}\kappa(s)\mathbf{n}(s) + o(h^2).\end{aligned}$$

26. PROPOSITION. Let M be an oriented $(n-1)$ -manifold, and let $f_1, f_2: M \rightarrow S^{n+}$ be two imbeddings such that $f_1(M)$ and $f_2(M)$ are convex and star-shaped with respect to $*$, and such that f_1 and f_2 induce the same metric on M , and the natural orientations on $f_1(M)$ and $f_2(M)$. Suppose, moreover, that the second fundamental forms of $f_1(M)$ and $f_2(M)$ are positive semi-definite. Then the same is true for the second fundamental forms of $\bar{f}_1(M)$ and $\bar{f}_2(M)$ in \mathbb{R}^n . (Note that under the given hypotheses, \bar{f}_1 and \bar{f}_2 will be immersions, by the Remark after Proposition 22.)

PROOF. Let c be an arclength parameterized curve in M (with the metric induced by f_1 or f_2). Apply (\ast) to the arclength parameterized curve $\gamma = f_1 \circ c$

in S^n , letting \mathbf{t}_1 and \mathbf{n}_1 be its tangent and normal, and κ_1 its curvature. We obtain

$$\begin{aligned} f_1(c(s+h)) &= \left(1 - \frac{h^2}{2}\right) f_1(c(s)) + h\mathbf{t}_1(s) + \frac{h^2}{2}\kappa_1(s)\mathbf{n}_1(s) + o(h^2) \\ &= \left(1 - \frac{h^2}{2}\right) x_1 + h\mathbf{t}_1 + \frac{h^2}{2}\kappa_1\mathbf{n}_1 + o(h^2) \quad \text{for short,} \end{aligned}$$

and similarly

$$\begin{aligned} f_2(c(s+h)) &= \left(1 - \frac{h^2}{2}\right) f_2(c(s)) + h\mathbf{t}_2(s) + \frac{h^2}{2}\kappa_2(s)\mathbf{n}_2(s) + o(h^2) \\ &= \left(1 - \frac{h^2}{2}\right) x_2 + h\mathbf{t}_2 + \frac{h^2}{2}\kappa_2\mathbf{n}_2 + o(h^2) \quad \text{for short.} \end{aligned}$$

So

$$\begin{aligned} (1) \quad \bar{f}_1(c(s+h)) &= \frac{\left(1 - \frac{h^2}{2}\right) x_1 + h\mathbf{t}_1 + \frac{h^2}{2}\kappa_1\mathbf{n}_1 + o(h^2) - \left\langle \left(1 - \frac{h^2}{2}\right) x_1 + h\mathbf{t}_1 + \frac{h^2}{2}\kappa_1\mathbf{n}_1 + o(h^2), * \right\rangle *}{\left\langle \left(1 - \frac{h^2}{2}\right) (x_1 + x_2) + h(\mathbf{t}_1 + \mathbf{t}_2) + \frac{h^2}{2}(\kappa_1\mathbf{n}_1 + \kappa_2\mathbf{n}_2) + o(h^2), * \right\rangle}. \end{aligned}$$

Writing this as

$$\frac{v - \frac{h^2}{2}v + hV}{\alpha - \frac{h^2}{2}\alpha + hA} \quad \begin{cases} v = x_1 - \langle x_1, * \rangle * \\ \alpha = \langle x_1 + x_2, * \rangle, \end{cases}$$

and noting that

$$\frac{v - \frac{h^2}{2}v + hV}{\alpha - \frac{h^2}{2}\alpha + hA} - \frac{v + hV}{\alpha + hA} = \frac{\frac{h^3}{2}(\alpha V - Av)}{\left(\alpha + \frac{h^2}{2}\alpha + hA\right)(\alpha + hA)},$$

we see that (1) can be written

$$(2) \quad \bar{f}_1(c(s+h)) = \frac{x_1 + h\mathbf{t}_1 + \frac{h^2}{2}\kappa_1\mathbf{n}_1 - \left\langle x_1 + h\mathbf{t}_1 + \frac{h^2}{2}\kappa_1\mathbf{n}_1, * \right\rangle *}{\left\langle x_1 + x_2 + h(\mathbf{t}_1 + \mathbf{t}_2) + \frac{h^2}{2}(\kappa_1\mathbf{n}_1 + \kappa_2\mathbf{n}_2), * \right\rangle} + O(h^3),$$

where $O(h^3)$ denotes a function such that $O(h^3)/h^3$ is bounded as $h \rightarrow 0$.

Expanding equation (2) out up to terms of order h^2 , we find that

$$\begin{aligned}
 (3) \quad & \bar{f}_1(c(s+h)) - \bar{f}_1(c(s)) \\
 &= \frac{h}{\langle x_1 + x_2, * \rangle} \left\{ -\langle \mathbf{t}_1 + \mathbf{t}_2, * \rangle \bar{f}_1(c(s)) + \mathbf{t}_1 - \langle \mathbf{t}_1, * \rangle * \right\} \\
 &+ \frac{h^2 \langle \mathbf{t}_1 + \mathbf{t}_2, * \rangle}{\langle x_1 + x_2, * \rangle^2} \left\{ \langle \mathbf{t}_1 + \mathbf{t}_2, * \rangle \bar{f}_1(c(s)) - \mathbf{t}_1 + \langle \mathbf{t}_1, * \rangle * \right\} \\
 &+ \frac{h^2}{2\langle x_1 + x_2, * \rangle} \left\{ -\langle \kappa_1 \mathbf{n}_1 + \kappa_2 \mathbf{n}_2, * \rangle \bar{f}_1(c(s)) + \kappa_1 \mathbf{n}_1 - \langle \kappa_1 \mathbf{n}_1, * \rangle * \right\} \\
 &+ O(h^3).
 \end{aligned}$$

Let \bar{N}_1 be the unit normal of $\bar{f}_1(M)$ at $\bar{f}_1(c(s))$. Clearly

$$\lim_{h \rightarrow 0} \left\langle \frac{\bar{f}_1(c(s+h)) - \bar{f}_1(c(s))}{h}, \bar{N}_1 \right\rangle = 0.$$

So equation (3) implies that

$$(4) \quad \left\langle -\langle \mathbf{t}_1 + \mathbf{t}_2, * \rangle \bar{f}_1(c(s)) + \mathbf{t}_1 - \langle \mathbf{t}_1, * \rangle *, \bar{N}_1 \right\rangle = 0.$$

Using (4), and the fact that $\langle \bar{N}_1, * \rangle = 0$ (since $\bar{f}_1(M)$ lies in $\mathbb{R}^n \times \{0\}$), equation (3) now gives

$$\begin{aligned}
 (5) \quad & \langle \bar{f}_1(c(s+h)) - \bar{f}_1(c(s)), \bar{N}_1 \rangle \\
 &= \frac{\kappa_1 h^2}{2\langle x_1 + x_2, * \rangle} \left\langle -\langle \mathbf{n}_1, * \rangle \bar{f}_1(c(s)) + \mathbf{n}_1, \bar{N}_1 \right\rangle \\
 &\quad - \frac{\kappa_2 h^2}{2\langle x_1 + x_2, * \rangle} \langle \mathbf{n}_2, * \rangle \langle \bar{f}_1(c(s)), \bar{N}_1 \rangle + O(h^3).
 \end{aligned}$$

Since Taylor's Theorem shows that the second derivative α'' of a function α is given by

$$\alpha''(x) = \lim_{h \rightarrow 0} \frac{\alpha(x+h) + \alpha(x-h) - 2\alpha(x)}{h^2},$$

equation (5) implies that

$$\begin{aligned}
 \langle (\bar{f}_1 \circ c)''(s), \bar{N}_1 \rangle &= \frac{\kappa_1}{\langle x_1 + x_2, * \rangle} \left\langle -\langle \mathbf{n}_1, * \rangle \bar{f}_1(c(s)) + \mathbf{n}_1, \bar{N}_1 \right\rangle \\
 &\quad - \frac{\kappa_2}{\langle x_1 + x_2, * \rangle} \langle \mathbf{n}_2, * \rangle \langle \bar{f}_1(c(s)), \bar{N}_1 \rangle.
 \end{aligned}$$

The term on the left is the second fundamental form of $\bar{f}_1(M)$ applied to $((\bar{f}_1 \circ c)'(s), (\bar{f}_1 \circ c)'(s))$. So it suffices to show that it is always ≥ 0 . Since

$$\frac{\kappa_1}{\langle x_1 + x_2, * \rangle} \geq 0 \quad \text{and} \quad \frac{\kappa_2}{\langle x_1 + x_2, * \rangle} \geq 0,$$

it suffices to show that

$$(6) \quad \left\langle -\langle \mathbf{n}_1, * \rangle \bar{f}_1(c(s)) + \mathbf{n}_1, \bar{N}_1 \right\rangle > 0 \quad \text{and} \quad -\langle \mathbf{n}_2, * \rangle \langle \bar{f}_1(c(s)), \bar{N}_1 \rangle > 0.$$

We can also assume that N_1 is the normal to the tangent plane of $f_1(M)$ at $f_1(c(s))$, since we can choose c so that $f_1 \circ c$ is a normal section of $f_1(M)$. Equation (6) is then proved in the following Lemma, whose statement introduces some more convenient notation.

27. LEMMA. Let P and Q be the tangent planes of $f_1(M)$ and $f_2(M)$ at the points $a_0 = f_1(p)$ and $b_0 = f_2(p)$, and let $c_0 = \bar{f}_1(p)$. Let a_n and b_n be the unit normals to P and Q at a_0 and b_0 , and let c_n be the unit normal to the tangent plane of $\bar{f}_1(M)$ at c_0 . Then

$$\left\langle -\langle a_n, * \rangle c_0 + a_n, c_n \right\rangle > 0 \quad \text{and} \quad -\langle b_n, * \rangle \cdot \langle c_0, c_n \rangle > 0.$$

PROOF. Choose positively oriented unit orthonormal vectors a_1, \dots, a_{n-1} at the point a_0 in P , let b_1, \dots, b_{n-1} be the corresponding vectors at b_0 in Q , and let c_1, \dots, c_{n-1} be the corresponding vectors at c_0 in the tangent plane of $\bar{f}_1(M)$ at c_0 . Then for some $C > 0$ we have

$$(7) \quad a_n = a_0 \times \cdots \times a_{n-1}, \quad b_n = b_0 \times \cdots \times b_{n-1}, \quad c_n = C \cdot * \times c_1 \times \cdots \times c_{n-1}.$$

Apply the formula for $d\bar{f}_1$, in the proof of Proposition 22, to the tangent vector X_i in M_p such that $df(X_i) = a_i$. This gives, in the present notation,

$$\langle a_0 + b_0, * \rangle^2 c_i = \langle a_0 + b_0, * \rangle [a_i - \langle a_1, * \rangle *] - \langle a_i + b_i, * \rangle [a_0 - \langle a_0, * \rangle *],$$

and thus

$$(8) \quad c_i = \frac{1}{\lambda_0^2} (\lambda_0 a_i - \lambda_i a_0) + (\cdots) * \quad i = 1, \dots, n-1,$$

where

$$(9) \quad \lambda_i = \langle a_i + b_i, * \rangle \quad i = 0, \dots, n-1.$$

Note also that c_0 is given by

$$(10) \quad c_0 = \frac{a_0}{\lambda_0} + (\cdots)*.$$

From (7) and (8) we obtain

$$\begin{aligned} (11) \quad c_n &= C \cdot * \times c_1 \times \cdots \times c_{n-1} \\ &= \frac{C}{\lambda_0^{2(n-1)}} \cdot * \times (\lambda_0 a_1 - \lambda_1 a_0) \times \cdots \times (\lambda_0 a_{n-1} - \lambda_{n-1} a_0) \\ &= \frac{C}{\lambda_0^n} \{ \lambda_0 (* \times a_1 \times \cdots \times a_{n-1}) - \lambda_1 (* \times a_0 \times a_2 \times \cdots \times a_{n-1}) \\ &\quad - \cdots - \lambda_{n-1} (* \times a_1 \times \cdots \times a_{n-2} \times a_0) \}. \end{aligned}$$

Consider first the quantity

$$-\langle b_n, * \rangle \cdot \langle c_0, c_n \rangle.$$

First of all, we have

$$\begin{aligned} (12) \quad \langle *, b_n \rangle &= \langle *, b_0 \times \cdots \times b_{n-1} \rangle \quad \text{by (7)} \\ &= \det \begin{pmatrix} b_0 \\ \vdots \\ b_{n-1} \\ * \end{pmatrix}. \end{aligned}$$

Also,

$$\begin{aligned} (13) \quad \langle c_0, c_n \rangle &= \left\langle \frac{a_0}{\lambda_0} + (\cdots)*, c_n \right\rangle \quad \text{by (10)} \\ &= \frac{C}{\lambda_0^n} \langle a_0, * \times a_1 \times \cdots \times a_{n-1} \rangle \quad \text{by (11)} \\ &= -\frac{C}{\lambda_0^n} \det \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \\ * \end{pmatrix}. \end{aligned}$$

Since f_1 and f_2 induce the natural orientations on $f_1(M)$ and $f_2(M)$, the determinants in (12) and (13) are both positive. Hence we do indeed have

$$-\langle b_n, * \rangle \cdot \langle c_0, c_n \rangle > 0.$$

Now consider

$$\langle -\langle a_n, * \rangle c_0 + a_n, c_n \rangle.$$

First of all, we have

$$(14) \quad -\langle a_n, * \rangle \cdot \langle c_0, c_n \rangle = \frac{C}{\lambda_0^n} \left[\det \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \\ * \end{pmatrix} \right]^2 \quad \text{by (7) and (13).}$$

Also,

$$\begin{aligned} \langle c_n, a_n \rangle &= \frac{C}{\lambda_0^n} \{ \lambda_0 \langle * \times a_1 \times \cdots \times a_{n-1}, a_0 \times \cdots \times a_{n-1} \rangle \\ &\quad - \lambda_1 \langle * \times a_0 \times \cdots \times a_{n-1}, a_0 \times \cdots \times a_{n-1} \rangle - \cdots \\ &\quad - \lambda_{n-1} \langle * \times a_1 \times \cdots \times a_0, a_0 \times \cdots \times a_{n-1} \rangle \} \quad \text{by (11).} \end{aligned}$$

Using the formula (Problem 5)

$$\langle v_1 \times \cdots \times v_n, w_1 \times \cdots \times w_n \rangle = \det(\langle v_i, w_j \rangle),$$

we obtain

$$\langle c_n, a_n \rangle = \frac{C}{\lambda_0^n} \{ \lambda_0 \langle *, a_0 \rangle + \lambda_1 \langle *, a_1 \rangle + \cdots + \lambda_{n-1} \langle *, a_{n-1} \rangle \}.$$

Substituting in from (8) yields

$$\begin{aligned} \langle c_n, a_n \rangle &= \frac{C}{\lambda_0^n} \cdot \sum_{i=0}^{n-1} \langle a_i, * \rangle^2 + \langle a_i, * \rangle \langle b_i, * \rangle \\ &\geq \frac{C}{2\lambda_0^n} \cdot \sum_{i=0}^{n-1} \langle a_i, * \rangle^2 - \langle b_i, * \rangle^2. \end{aligned}$$

Since

$$\sum_{i=0}^n \langle a_i, * \rangle^2 = 1 = \sum_{i=0}^n \langle b_i, * \rangle^2,$$

we get

$$\begin{aligned} (15) \quad \langle c_n, a_n \rangle &\geq \frac{C}{2\lambda_0^n} (\langle *, b_n \rangle^2 - \langle *, a_n \rangle^2) \\ &= \frac{C}{2\lambda_0^n} \left[\left[\det \begin{pmatrix} b_0 \\ \vdots \\ b_{n-1} \\ * \end{pmatrix} \right]^2 - \left[\det \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \\ * \end{pmatrix} \right]^2 \right]. \end{aligned}$$

From (14) and (15) we get

$$\langle -\langle a_n, * \rangle c_0 + a_n, c_n \rangle \geq \frac{C}{2\lambda_0^n} \left[\left[\det \begin{pmatrix} b_0 \\ \vdots \\ b_{n-1} \\ * \end{pmatrix} \right]^2 + \left[\det \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \\ * \end{pmatrix} \right]^2 \right] > 0. \quad \blacklozenge$$

There is a result analogous to Proposition 26 when we begin with imbeddings into \mathbb{R}^n and construct the imbeddings into S^n , but we will not need it. It is probably already clear how the results which we have just proved can be used to transfer theorems from Euclidean space to the sphere. For example, keeping to dimension 3, which is the really interesting one, suppose we have two compact convex surfaces $M, \bar{M} \subset S^3$ (each contained in some open hemisphere), and an isometry $\alpha: M \rightarrow \bar{M}$. We claim that α is the restriction of an isometry $A: S^3 \rightarrow S^3$. Without loss of generality, we can assume that M and \bar{M} are contained in S^{n+} and are star-shaped with respect to $*$. Let $f_1: M \rightarrow S^{n+}$ be the inclusion map, and let $f_2: M \rightarrow S^{n+}$ be $\alpha \circ f_1$; then f_1 and f_2 induce the same metric on M . We can also assume that M is oriented so that f_1 and f_2 induce the natural orientation on M and \bar{M} , by composing f_2 with a reflection if necessary. Then $\bar{f}_1, \bar{f}_2: M \rightarrow \mathbb{R}^3$ induce the same metric on M , by Proposition 22, and $\bar{f}_1(M)$ and $\bar{f}_2(M)$ have $K \geq 0$ by Proposition 26. So by Theorem 12, there is an isometry $B: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with $\bar{f}_2 = B \circ \bar{f}_1$. Then Proposition 25 shows that there is an isometry $A: S^3 \rightarrow S^3$ with

$$f_2 = \hat{f}_2 = A \circ \hat{f}_1 = A \circ f_1,$$

which shows that

$$\alpha \circ f_1 = A \circ f_1 \implies \alpha \text{ is the restriction of } A \text{ to } f_1(M) = M.$$

Pogorelov states that essentially the same formulas can be used to transfer rigidity problems from hyperbolic space to Euclidean space, and he shows how problems of infinitesimal rigidity can also be transferred in this way.

We also want to add a few remarks, of a different sort, about hypersurfaces of S^{n+1} and H^{n+1} . The proof of Theorem 1 carries over almost without change to this situation, so hypersurfaces of S^{n+1} or H^{n+1} are rigid if their type number (the rank of $X_p \mapsto \nabla'_{X_p} \nu$) is ≥ 3 at each point p . The hypersurfaces of S^{n+1} and H^{n+1} with type number 2 at all points were studied by Dolbeault-Lemoine [1]. She divides them into the same three classes that É. Cartan found

for hypersurfaces of \mathbb{R}^{n+1} , but it turns out that all hypersurfaces in one of the classes are rigid in S^{n+1} and H^{n+1} for any $n \geq 3$, while the hypersurfaces of the other two classes are rigid in S^{n+1} and H^{n+1} for any $n \geq 4$. Moreover, the hypersurfaces with type number 1 at all points are also rigid in S^{n+1} and H^{n+1} for any $n \geq 4$. This leads her to conclude that for $n \geq 4$, all hypersurfaces of S^{n+1} and H^{n+1} are rigid. Unfortunately, this does not follow directly from the preceding results, for it is conceivable that two hypersurfaces from different classes can be joined together in two isometric, but non-congruent, ways; whether this is actually possible is a question which still has to be cleared up.

The only subject left for us to consider at this point is the rigidity of submanifolds of higher codimension. There are two main results in this direction, a classical local one, and a modern global one.

The classical result involves the notion of the type number of a submanifold $M^n \subset \mathbb{R}^m$ of arbitrary codimension. First an algebraic definition. Let V be a vector space, and let $T_1, \dots, T_k: V \rightarrow V$ be linearly independent linear transformations. We define the **type number** of $\{T_1, \dots, T_k\}$ to be the largest integer t for which there are t vectors $v_1, \dots, v_t \in V$ such that the kt vectors

$$T_r(v_i) \quad 1 \leq i \leq t, \quad 1 \leq r \leq k$$

are linearly independent. The **type number** of linearly independent matrices S_1, \dots, S_k is defined as that of the corresponding linear transformations. Now for a point $p \in M^n \subset \mathbb{R}^m$ we let $k(p)$ be the rank of the map $\xi \mapsto A_\xi$ from M_p^\perp into the space of all symmetric maps of M_p into itself (recall that $A_\xi: M_p \rightarrow M_p$ is defined by $\langle A_\xi(X), Y \rangle = \langle s(X, Y), \xi \rangle$). Equivalently, $k(p)$ is the dimension of the **first normal space** at p , which may be defined as the orthogonal complement in M_p^\perp of $\{\xi: A_\xi = 0\}$ (compare Addendum 4 of Chapter 7). Set $k = k(p)$ and let ξ_1, \dots, ξ_k be a basis for the first normal space at p . If $T_r: M_p \rightarrow M_p$ is A_{ξ_r} for $r = 1, \dots, k$, then T_1, \dots, T_k are linearly independent, and we can define the **type number** $t(p)$ of M at p to be the type number of $\{T_1, \dots, T_k\}$; it is easily checked that this definition does not depend on the choice of ξ_1, \dots, ξ_k . The following result shows that submanifolds with type number at least 2 cannot twist too much.

28. LEMMA. Let N^m be a manifold of constant curvature, and let M^n be a submanifold with normal connection D , whose first normal space $\text{Nor}^1 M_p$ has the same dimension k at all points, and whose type number is ≥ 2 at all points. Then $D_Z \xi \in \text{Nor}^1 M_p$ for any section ξ of $\text{Nor}^1 M_p$ and $Z \in M_p$.

Consequently, if M is connected, then it lies in some $(n + k)$ -dimensional totally geodesic subspace of N .

PROOF. Locally we can choose orthonormal sections v_{n+1}, \dots, v_m of $\text{Nor } M$ such that v_{n+1}, \dots, v_{n+k} span $\text{Nor}^1 M$. Then $A_{v_r} = 0 \implies \text{II}^r = 0$ for $r > n + k$. So the Codazzi-Mainardi equations (Theorem 7-14) give

$$\begin{aligned}
 0 &= \sum_{s=n+1}^{n+k} \text{II}^s(Y, W) \beta_s^r(X) - \text{II}^s(X, W) \beta_s^r(Y) \quad r > n + k \\
 &\Downarrow \\
 (1) \quad 0 &= \sum_{s=n+1}^{n+k} \beta_s^r(X) \cdot A_{v_s}(Y) - \beta_s^r(Y) \cdot A_{v_s}(X) \quad r > n + k.
 \end{aligned}$$

By assumption, there are $X, Y \in M_p$ such that the vectors $A_{v_s}(X), A_{v_s}(Y)$ for $s = n + 1, \dots, n + k$ are linearly independent. Then (1) shows that

$$(2) \quad \beta_s^r(X) = \beta_s^r(Y) = 0 \quad n + 1 \leq s \leq n + k, \quad r > n + k.$$

Moreover, for any $Z \in M_p$ and $r > n + k$ we have

$$\begin{aligned}
 0 &= \sum_{s=n+1}^{n+k} \beta_s^r(X) \cdot A_{v_s}(Z) - \beta_s^r(Z) \cdot A_{v_s}(X) \\
 &= - \sum_{s=n+1}^{n+k} \beta_s^r(Z) \cdot A_{v_s}(X) \quad \text{by (2)} \\
 &\Downarrow \\
 (3) \quad \beta_s^r(Z) &= 0 \quad n + 1 \leq s \leq n + k, \quad r > n + k.
 \end{aligned}$$

This shows that $D_Z v^s \in \text{Nor}^1 M_p$ for $n + 1 \leq s \leq n + k$, and proves the first part of the theorem.

Now consider the $(n + k)$ -dimensional distribution $\Delta(p) = M_p \oplus \text{Nor}^1 M_p$ along M . The first part of the theorem clearly implies that $\nabla'_Z \xi \in \Delta(p)$ for all sections ξ of Δ and $Z \in M_p$. So Δ is parallel along any curve c , by Pre-Lemma 7-7. The result then follows from Corollary 7-11. ♦

Remark: A curve in \mathbb{R}^m with $\kappa_1, \dots, \kappa_m$ all non-zero represents a counterexample to Lemma 28 when the type number is < 2 .

The extension of Theorem 1 of this chapter to submanifolds of higher codimension rests on some more algebraic results.

29. LEMMETTE. Let $\phi_r, \psi_r, \bar{\phi}_r, \bar{\psi}_r \in V^*$ for $r = 1, \dots, k$ with

$$\sum_{r=1}^k \phi_r \wedge \psi_r = \sum_{r=1}^k \bar{\phi}_r \wedge \bar{\psi}_r.$$

Suppose that $\phi_1, \dots, \phi_k, \psi_1, \dots, \psi_k$ are linearly independent. Then the same is true of the $\bar{\phi}_r, \bar{\psi}_r$, and the subspace $[\phi_1, \dots, \phi_k, \psi_1, \dots, \psi_k]$ spanned by the ϕ_r and ψ_r equals the subspace $[\bar{\phi}_1, \dots, \bar{\phi}_k, \bar{\psi}_1, \dots, \bar{\psi}_k]$ spanned by the $\bar{\phi}_r$ and $\bar{\psi}_r$.

PROOF. Recall that for $v \in V$ and $\omega \in \Omega^k(V)$ we define $v \lrcorner \omega \in \Omega^{k-1}(V)$ by $v \lrcorner \omega(v_1, \dots, v_{k-1}) = \omega(v, v_1, \dots, v_{k-1})$. Define a map $f: V \rightarrow V^*$ by

$$f(v) = v \lrcorner \left(\sum_{r=1}^k \phi_r \wedge \psi_r \right) = \sum_{r=1}^k \phi_r(v) \cdot \psi_r - \psi_r(v) \cdot \phi_r.$$

Clearly

$$\text{range } f \subset [\phi_1, \dots, \phi_k, \psi_1, \dots, \psi_k].$$

Moreover, by linear independence of the ϕ_r and ψ_r , there is $v \in V$ with $\phi_1(v) = 1$ and all other $\phi_r(v) = \psi_r(v) = 0$. Then $f(v) = \psi_1$, so $\psi_1 \in \text{range } f$. Similarly, all $\phi_r, \psi_r \in \text{range } f$, and we therefore have

$$\text{range } f = [\phi_1, \dots, \phi_k, \psi_1, \dots, \psi_k].$$

Now we also have

$$f(v) = \sum_{r=1}^k \bar{\phi}_r(v) \cdot \bar{\psi}_r - \bar{\psi}_r(v) \cdot \bar{\phi}_r.$$

So

$$[\phi_1, \dots, \phi_k, \psi_1, \dots, \psi_k] = \text{range } f \subset [\bar{\phi}_1, \dots, \bar{\phi}_k, \bar{\psi}_1, \dots, \bar{\psi}_k].$$

Hence the subspace on the right has dimension $\geq 2k$. So it has dimension exactly $2k$, which means that the $\bar{\phi}_r, \bar{\psi}_r$ are linearly independent and that the two subspaces are equal. ♦

30. LEMMA (CHERN). Let $S_1, \dots, S_k, \bar{S}_1, \dots, \bar{S}_k$ be symmetric $n \times n$ matrices, with S_1, \dots, S_k linearly independent, of type number ≥ 3 . Suppose that the sum of the determinants of corresponding 2×2 submatrices of the S_r always equals the sum of the determinants of the corresponding 2×2 submatrices of the \bar{S}_r . Then we have

$$\bar{S}_r = \sum_{s=1}^k A_{sr} S_s$$

for some *orthogonal* matrix $A \in O(k)$.

PROOF. Let $T_r, \bar{T}_r: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the linear transformations with matrices S_r, \bar{S}_r . The hypothesis on the determinants of S_r, \bar{S}_r is equivalent to the hypothesis that the maps

$$T_r^*, \bar{T}_r^*: \Omega^2(\mathbb{R}^n) \rightarrow \Omega^2(\mathbb{R}^n)$$

satisfy

$$\sum_{r=1}^k T_r^* = \sum_{r=1}^k \bar{T}_r^*;$$

in other words, for all $\phi_i, \phi_j \in V^*$ we have

$$(1) \quad \sum_{r=1}^k T_r^*(\phi_i) \wedge T_r^*(\phi_j) = \sum_{r=1}^k \bar{T}_r^*(\phi_i) \wedge \bar{T}_r^*(\phi_j).$$

Choose a basis $\{\phi_i\}$ for V^* such that the $3k$ vectors $\{T_r^*(\phi_i) : i = 1, 2, 3\}$ are linearly independent. Applying (1) with $i = 1, j = 2$, and using the Lemmette, we see that each $\bar{T}_r^*(\phi_1)$ is a linear combination of the $T_s^*(\phi_1), T_s^*(\phi_2)$. Similarly, each $\bar{T}_r^*(\phi_1)$ is a linear combination of the $T_s^*(\phi_1), T_s^*(\phi_3)$. So each $\bar{T}_r^*(\phi_1)$ is a linear combination of the $T_s^*(\phi_1)$. The analogous conclusions hold for the $\bar{T}_r^*(\phi_2)$ and the $\bar{T}_r^*(\phi_3)$. Set

$$\bar{T}_r^*(\phi_1) = \sum_s B_{rs} T_s^*(\phi_1)$$

$$\bar{T}_r^*(\phi_2) = \sum_s C_{rs} T_s^*(\phi_2)$$

$$\bar{T}_r^*(\phi_3) = \sum_s D_{rs} T_s^*(\phi_3).$$

Equation (1) gives us

$$\begin{cases} B \cdot C^t = I = C \cdot B^t & (i = 1, j = 2 \text{ and } i = 2, j = 1) \\ C \cdot D^t = I = D \cdot C^t & (i = 2, j = 3 \text{ and } i = 3, j = 2) \\ B \cdot D^t = I = D \cdot B^t & (i = 1, j = 3 \text{ and } i = 3, j = 1). \end{cases}$$

These imply that $B = C = D$ and $B \cdot B^t = I$, so that B is orthogonal. Let $\tilde{T}_r: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the linear transformation with

$$\tilde{T}_r^* = \sum_s B_{sr} T_s^*.$$

Then $\tilde{T}_r^*(\phi_j) = \bar{T}_r^*(\phi_j)$ for $j = 1, 2, 3$. We just have to show that this is also true for $j \geq 4$.

Using (1) and orthogonality of B , we see that

$$\sum_r \bar{T}_r^*(\phi_j) \wedge \bar{T}_r^*(\phi_i) = \sum_r \tilde{T}_r^*(\phi_j) \wedge \tilde{T}_r^*(\phi_i) \quad i = 1, 2, 3$$

and hence

$$\sum_r \{\bar{T}_r^*(\phi_j) - \tilde{T}_r^*(\phi_j)\} \wedge \bar{T}_r^*(\phi_i) = 0 \quad i = 1, 2, 3.$$

The $\bar{T}_r^*(\phi_i)$ are linearly independent, so by Cartan's Lemma (Lemma 1-13 or Problem I.7-11) the $\bar{T}_r^*(\phi_j) - \tilde{T}_r^*(\phi_j)$ are a linear combination of the $\bar{T}_r^*(\phi_i)$ for each $i = 1, 2$. But $\{\bar{T}_r^*(\phi_i) : i = 1, 2\}$ are linearly independent, since $\{T_r^*(\phi_i) : i = 1, 2\}$ are, so we must have $\bar{T}_r^*(\phi_j) - \tilde{T}_r^*(\phi_j) = 0$. ♦

31. THEOREM (ALLENDOERFER). Let M^n and \bar{M}^n be immersed submanifolds of \mathbb{R}^m , and let $\phi: M \rightarrow \bar{M}$ be an isometry. Suppose that the first normal spaces of M and \bar{M} have the same constant dimension k at all points. Suppose, moreover, that the type number is ≥ 3 at all points of M . Then ϕ is the restriction of a Euclidean motion.

PROOF. By Lemma 28, there is no loss of generality in assuming that $m = n+k$, so that the first normal space is the whole normal space. First we will show that ϕ is locally the restriction of a Euclidean motion. Choose an orthonormal moving frame X_1, \dots, X_n in a neighborhood U of p , and let $\bar{X}_i = \phi_* X_i$. Choose orthonormal sections v_{n+1}, \dots, v_m of the normal bundle of M , and $\bar{v}_{n+1}, \dots, \bar{v}_m$ for the normal bundle of \bar{M} . For $q \in U$, define $n \times n$ symmetric matrices S_r , $r = n+1, \dots, m$ by

$$(S_r)_{ij} = \Pi^r(X_i(q), X_j(q)).$$

Define \bar{S}_r similarly, for the point $\phi(q)$. Gauss' equation, in the form given on pg. IV.32, shows that the S_r, \bar{S}_r satisfy the hypotheses of Lemma 30. Thus we see that there is an orthogonal matrix-valued function A on U with

$$\bar{S}_r = \sum_s A_{sr} S_s.$$

Using linear independence of the S_r , and smoothness of the S_r and \bar{S}_r , we see that the A_{sr} vary smoothly with q . Now define new sections $v'_{n+1}, \dots, v'_{n+k}$ of the normal bundle of M by

$$v'_r = \sum_s A_{sr} v_s.$$

Then for the corresponding second fundamental forms we have

$$\begin{aligned}\Pi'^r(X, Y) &= \langle s(X, Y), v'_r \rangle \\ &= \left\langle s(X, Y), \sum_s A_{sr} v_s \right\rangle \\ &= \sum_s A_{sr} \Pi^s.\end{aligned}$$

In particular,

$$\begin{aligned}\Pi'^r(X_i, X_j) &= \sum_s A_{sr} \Pi^s(X_i, X_j) \\ &= \sum_s A_{sr} (S_s)_{ij} \\ &= (\bar{S}_r)_{ij} \\ &= \bar{\Pi}^r(\bar{X}_i, \bar{X}_j).\end{aligned}$$

Theorem 7-19 then shows that ϕ is the restriction of a Euclidean motion on U .

We claim that this Euclidean motion is unique. This is easy to see once we note that since the first normal space is the whole normal space, every normal vector at $p \in M$ is $c''(0)$ for some arclength parameterized curve in M . Having established uniqueness, it is clear that the local result implies the global one. ♦

To be sure, the hypothesis that the type number is ≥ 3 is extremely strong for submanifolds of higher codimension, but Theorem 31 is very likely the best local result obtainable. It is therefore a pleasant surprise to find that there is a global result in this area. To end this chapter, with its vast areas of ignorance, on a more joyful note, we quote the following beautiful recent result of J. C. Moore; the proof is somewhat lengthy, but uses only material which has already been developed here, the only somewhat non-standard result being Corollary 11-6.

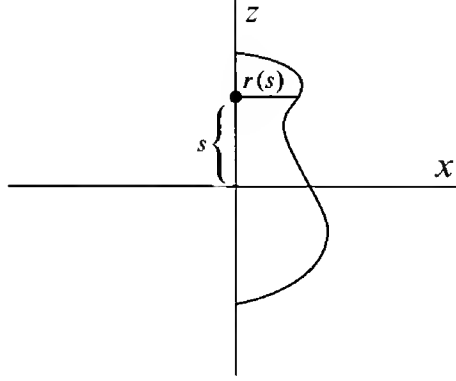
THEOREM (J. C. MOORE). If M_1, \dots, M_k are compact connected Riemannian manifolds with M_i of dimension $n_i \geq 2$, then any immersion $\phi: M_1 \times \dots \times M_k \rightarrow \mathbb{R}^{n_1 + \dots + n_k + k}$ is, up to a Euclidean motion, a product of immersions of the M_i as hypersurfaces.

In particular, if the M_i are compact convex surfaces in \mathbb{R}^3 , then $M_1 \times \dots \times M_k$ is rigid in \mathbb{R}^{3k} , though not locally rigid.

ADDENDUM

INFINITESIMAL BENDINGS OF ROTATION SURFACES

We will be dealing with surfaces of revolution obtained by revolving a curve $c(s) = (r(s), s)$ in the (x, z) -plane around the z -axis. Define



$$\gamma(t) = (\cos t, \sin t, 0),$$

so that in time 2π , the curve γ goes once around a unit circle in the (x, y) -plane. Then our rotation surface is given by

$$f(s, t) = r(s) \cdot \gamma(t) + s \cdot e_3,$$

where $e_3 = (0, 0, 1)$. Here $r(s)$ is smooth for $s_0 < s < s_1$, while $r'(s_0) = -\infty$ and $r'(s_1) = +\infty$. Considering for the moment only $s \in (s_0, s_1)$, any vector field Z along f can be written uniquely as a linear combination of the three vectors γ, γ', e_3 , thus

$$Z(s, t) = a(s, t) \cdot \gamma(t) + b(s, t) \cdot \gamma'(t) + c(s, t) \cdot e_3.$$

Now Z is an infinitesimal bending if and only if

$$(1) \quad \langle f_1, Z_1 \rangle = 0, \quad \langle f_2, Z_2 \rangle = 0, \quad \langle f_1, Z_2 \rangle + \langle f_2, Z_1 \rangle = 0.$$

Since

$$\begin{aligned} f_1 &= r' \cdot \gamma + e_3 & f_2 &= r \cdot \gamma' \\ Z_1 &= a_1 \cdot \gamma + b_1 \cdot \gamma' + c_1 e_3 & Z_2 &= (a_2 - b) \gamma + (b_2 + a) \gamma' + c_2 e_3, \end{aligned}$$

equations (1) become

$$\begin{aligned} (2) \quad & r' a_1 + c_1 = 0 \\ & b_2 + a = 0 \\ & r'(a_2 - b) + r b_1 + c_2 = 0. \end{aligned}$$

Since a, b, c must be periodic in t of period 2π , it is natural to look for solutions in terms of Fourier series

$$a(s, t) = \sum_{k=-\infty}^{\infty} e^{ikt} \phi_k(s)$$

$$b(s, t) = \sum_{k=-\infty}^{\infty} e^{ikt} \psi_k(s)$$

$$c(s, t) = \sum_{k=-\infty}^{\infty} e^{ikt} \xi_k(s),$$

where, in order that ϕ_k, ψ_k, ξ_k should be real-valued, we must have

$$(3) \quad \phi_{-k} = \bar{\phi}_k, \quad \psi_{-k} = \bar{\psi}_k, \quad \xi_{-k} = \bar{\xi}_k.$$

For a (complex-valued) solution involving a single k , equations (2) become

$$(4) \quad \begin{aligned} r'(s)\phi_k'(s) + \xi_k'(s) &= 0 \\ ik\psi_k(s) + \phi_k(s) &= 0 \\ r'(s)[ik\phi_k(s) - \psi_k(s)] + r(s)\psi_k'(s) + ik\xi_k(s) &= 0. \end{aligned}$$

Differentiating the third equation, and then using the first two, we obtain

$$(*) \quad r\psi_k'' + (k^2 - 1)r''\psi_k = 0.$$

Conversely, suppose we have a complex-valued function ψ_k satisfying (*). Then the first two equations of (4) can be used to determine ϕ_k and then ξ_k . If we *define* $\phi_{-k}, \psi_{-k}, \xi_{-k}$ by (3), then (4) also holds for $-k$. Thus we will have real-valued solutions

$$\begin{aligned} a(s, t) &= e^{ikt} \phi_k(s) + e^{-ikt} \bar{\phi}_k(s) \\ b(s, t) &= e^{ikt} \psi_k(s) + e^{-ikt} \bar{\psi}_k(s) \\ c(s, t) &= e^{ikt} \xi_k(s) + e^{-ikt} \bar{\xi}_k(s) \end{aligned}$$

of (2). Any finite linear combination of solutions is also a solution.

For $k = 0$, we can solve directly from (4). The first two equations give $\phi_0 = 0$, and then $\xi_0 = \text{constant } A$. Then the third equation gives

$$0 = -r'\psi_0 + r\psi_0' = r^2 \left(\frac{\psi_0}{r} \right)' \implies \psi_0 = Br \quad \text{for some constant } B.$$

Thus we obtain the infinitesimal bending

$$\begin{aligned} Z_0(s, t) &= Br(s) \cdot \gamma'(t) + A \cdot e_3 \\ &= [r(s) \cdot \gamma(t) + s \cdot e_3] \times Be_3 + Ae_3 \\ &= (f(s, t) \times Be_3) + Ae_3, \end{aligned}$$

which is trivial.

For $k = 1$, equation (*) says that ψ_1 is linear. In particular, one possible solution is

$$\psi_1(s) = Cs, \quad C \text{ real.}$$

Then the second equation of (4) gives

$$\phi_1(s) = -iCs,$$

and so the first equation gives $\xi_1(s) = iCr(s) + \text{constant}$; in particular, we can take

$$\xi_1(s) = iCr(s).$$

Then we have

$$\begin{aligned} a(s, t) &= 2 \operatorname{Re} e^{it} (-iCs) = 2Cs \sin t \\ b(s, t) &= 2 \operatorname{Re} e^{it} Cs = 2Cs \cos t \\ c(s, t) &= 2 \operatorname{Re} e^{it} (iCr(s)) = -2Cr(s) \sin t. \end{aligned}$$

This gives the trivial infinitesimal bending

$$\begin{aligned} \frac{1}{2} Z_1(s, t) &= (Cs \sin t) \gamma(t) + (Cs \cos t) \gamma'(t) - (Cr(s) \sin t) e_3 \\ &= [r(s) \cdot \gamma(t) + se_3] \times [-C \sin t \gamma' + C \cos t \gamma] \\ &= f(s, t) \times (C, 0, 0). \end{aligned}$$

Similarly, we can take

$$\psi_1(s) = iCs \quad (C \text{ real}), \quad \phi_1(s) = Cs, \quad \xi_1(s) = -Cr(s),$$

obtaining the infinitesimal bending

$$\begin{aligned} \frac{1}{2} Z_1(s, t) &= [r(s) \cdot \gamma(t) + se_3] \times [-C \cos t \gamma' - C \sin t \gamma] \\ &= f(s, t) \times (0, -C, 0). \end{aligned}$$

Obviously *every* trivial infinitesimal bending is a linear combination of the various infinitesimal bendings Z_0, Z_1 . Since the various Z_k are linearly independent, we see that any solution of (*) for $k \geq 2$ leads to an infinitesimal bending Z which is *not* trivial.

These considerations all hold only for the region $s_0 < s < s_1$, that is, for the surface of revolution minus its two “poles”. Given any infinitesimal bending Z_k obtained as above, we still have to see how it behaves at the poles (one can easily see, for example, that if we had picked $\psi_1(s) = Cs + D$ with $D \neq 0$, then the corresponding solution, although not a trivial infinitesimal bending, would have a singularity at the poles). In order to do this, we consider the functions $\phi_k \circ r^{-1}$, etc., where r^{-1} really denotes two different functions, depending on which pole we are at. To be precise, we note that at either pole $r^{-1}(x)$ makes sense for x in some interval $[0, \varepsilon)$. We extend r^{-1} to a function ρ on $(-\varepsilon, \varepsilon)$ by requiring ρ to be even. We will assume that ρ is analytic at 0 (this is precisely what one needs in order for the surface f to be analytic at the poles). Setting

$$\Phi_k = \phi_k \circ \rho$$

$$\Psi_k = \psi_k \circ \rho$$

$$\Xi_k = \xi_k \circ \rho,$$

and noting that $\rho' = 1/r' \circ \rho$, we find that equations (4) can be written

$$\begin{aligned} \rho'(x)\Xi_k'(x) + \Phi_k'(x) &= 0 \\ (5) \quad ik\Psi_k(x) + \Phi_k(x) &= 0 \\ ik\Phi_k(x) - \Psi_k(x) + x\Psi_k'(x) + ik\rho'(x)\Xi(x) &= 0, \end{aligned}$$

while equation (*) becomes

$$(**) \quad x\rho'(x)\Psi_k''(x) - x\rho''(x)\Psi_k'(x) - (k^2 - 1)\rho''(x)\Psi_k(x) = 0.$$

Now suppose also that $\rho''(0) \neq 0$. Then

$$\rho'(x) = \rho''(0)x[1 + \cdots], \quad \rho''(x) = \rho''(0) \cdot [1 + ***]$$

and we can write our equation as

$$x^2\rho''(0)[1 + \cdots]\Psi_k''(x) - x\rho''(0)[1 + \cdots]\Psi_k'(x) - (k^2 - 1)\rho''(0)[1 + ***]\Psi_k(x) = 0$$

or

$$(***) \quad x^2\Psi_k''(x) + x\alpha(x)\Psi_k'(x) + \beta(x)\Psi_k(x) = 0,$$

where α and β are analytic with

$$\alpha(0) = -1, \quad \beta(0) = 1 - k^2.$$

Now equation (***) has a “singular point” at 0; we cannot put it in the form $\Psi_k''(x) = F(x, \Psi_k(x), \Psi_k'(x))$ near 0, so we cannot apply our standard theorems. However, this singular point is of a very special sort, called a “regular singular point”, and there is a complete theory to cover this situation. It is one of the standard topics in differential equations, which the reader can find, for example, in Whittaker and Watson [1; Chapter 10]. The theory shows that equation (***), or equivalently (**), has two linearly independent solutions near $x = 0$ of the form

$$\begin{aligned} \mu_1(x) &= x^{1+k} \cdot (\text{analytic function}) \\ \mu_2(x) &= x^{1-k} \cdot (\text{analytic function}) + c(\log x) \cdot \mu_1(x), \end{aligned}$$

where the analytic functions in question are non-zero at 0 (but c might be 0). So when $k \geq 2$, we see that (**) always has one solution which is analytic near 0, while any linearly independent solution blows up at zero. If Ψ_k is an analytic solution, so that Ψ_k vanishes up to order $k + 1$ at 0, then the first and second equations of (5) determine functions Φ_k and Ξ_k which vanish up to order $k + 1$ and k , respectively, at 0. So for $k \geq 2$ the infinitesimal bending Z_k then determined by $\phi_k = \Phi_k \circ r$, etc., has the form

$$Z_k(s, t) = r(s)^2 \sigma(r(s), t)$$

for some analytic function σ . Now an analytic parameterization of our rotation surface near a pole is given by

$$\begin{aligned} (x, y) &\mapsto \left(x, y, r^{-1}(\sqrt{x^2 + y^2}) \right) \\ &= f \left(r^{-1}(\sqrt{x^2 + y^2}), \arctan \frac{y}{x} \right). \end{aligned}$$

So if $\tilde{Z}_k(x, y)$ denotes the value of Z_k at this point on the rotation surface, then

$$\tilde{Z}_k(x, y) = (\sqrt{x^2 + y^2})^2 \sigma(\sqrt{x^2 + y^2}, t) = (x^2 + y^2) \sigma(\sqrt{x^2 + y^2}, t).$$

One can easily check that this function is* C^2 at 0.

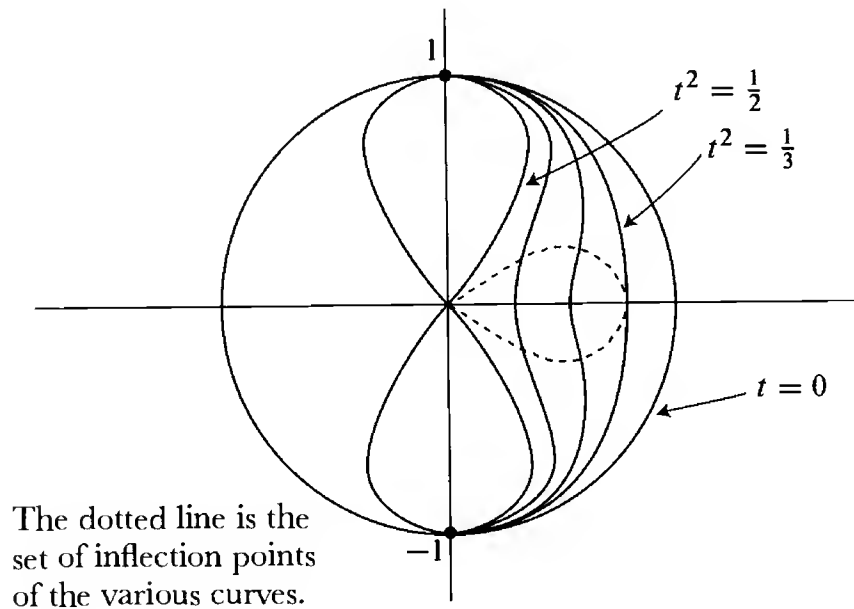
* If σ involved only even powers of $\sqrt{x^2 + y^2}$, then we would actually have an analytic function. Unfortunately, this can never happen, since equations (5), and the fact that σ is even, shows that Φ, Ψ, Ξ cannot all be even functions.

This analysis holds at each pole, but it will usually happen that the function Ψ_k which is analytic for one choice of $\rho = r^{-1}$ will not be analytic for the other choice. So further analysis is required.

We consider a 1-parameter family of functions r_t which passes continuously from a convex function to non-convex functions. For example, we can determine r_t by the equation

$$(r_t(s)^2 + s^2)^2 + 2t^2(r_t(s)^2 - s^2) = 1 - 2t^2 \quad 0 \leq t^2 < \frac{1}{2}.$$

For $t^2 = 1/2$ we have a lemniscate, and for $t = 0$ we have a circle. For $t^2 \leq 1/3$ the functions are convex.



For a fixed t , which we temporarily suppress, consider equation (*) for $r = r_t$:

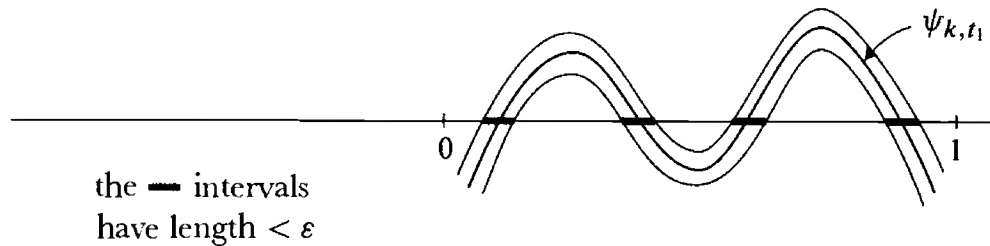
$$(*) \quad r\psi_k'' + (k^2 - 1)r''\psi_k = 0.$$

On any (concave) interval where $r'' > 0$ (which exist for $t^2 > 1/3$), there will be a large number of zeros of ψ_k once k is large enough. To prove this we merely choose a subinterval on which $r''/r > \varepsilon > 0$ for some ε , and apply the Sturm Comparison Theorem (Theorem 9-15) to equation (*) and the equation $y'' + (k^2 - 1)\varepsilon y = 0$, which has lots of zeros for large k . Notice that on each of the outer intervals, where $r'' < 0$, the function ψ_k cannot have a positive maximum ($\psi_k > 0$, $\psi_k'' \leq 0$) or a negative minimum, so it cannot have even two zeros. Thus the total number of zeros in $(-1, 1)$ is finite. For $t^2 > 1/3$ we have $r'' < 0$ everywhere on $(-1, 1)$, so ψ_k has at most one zero on $(-1, 1)$. Even for $t^2 = 1/3$ we easily see that ψ_k has at most 2 zeros on $(-1, 1)$. Since ψ_k

is a solution of a second order equation, we must have $\psi_k' \neq 0$ when $\psi_k = 0$ (assuming ψ_k is not the zero function), so ψ_k crosses the axis at each zero, rather than being tangent to it. It follows easily that any function sufficiently close to ψ_k has at least as many zeros as ψ_k .

Now pick some t_0 with $t_0^2 > 1/3$ and a k so large that any solution ψ_{k,t_0} of (*) for $r = r_{t_0}$ has more than 2 zeros on $[0, 1)$. Let ψ_{k,t_0} be a solution which gives an infinitesimal bending that is C^2 at the bottom pole $s = -1$, and let $n_0 > 2$ be the number of its zeros on $[0, 1)$. Now for all $t < t_0$ we will pick a continuous family of solutions $\psi_{k,t}$ of (*) for $r = r_t$, all of which also give infinitesimal bendings that are C^2 at the bottom pole. Choosing $\psi_{k,t}$ is equivalent to choosing $\Psi_{k,t}$, which is determined only up to a constant factor. So we can easily arrange that each $\psi_{k,t}$ is not the zero function. For $t^2 \geq 1/3$, the function $\psi_{k,t}$ has at most two zeros in $(-1, 1)$, so *a fortiori* at most two zeros in $[0, 1)$. So we can consider the greatest lower bound t_1 of all t such that $\psi_{k,t}$ has exactly n_0 zeros on $[0, 1)$, and $t_1^2 > 1/3$.

We claim that ψ_{k,t_1} has a zero at 0. For suppose that all the zeros of ψ_{k,t_1} on $[0, 1)$ actually occur on $(0, 1)$, and let n_1 be the number of such zeros. Since any function sufficiently close to ψ_{k,t_1} has at least as many zeros on $(0, 1)$ as ψ_{k,t_1} does, we easily see that $n_1 \leq n_0$. We claim that we cannot have $n_1 < n_0$. To prove this, it suffices to show that the functions $\psi_{k,t}$ for t close to t_1 have *at most* as many zeros on $(0, 1)$ as ψ_{k,t_1} does. On some interval containing the n_1 zeros of ψ_{k,t_1} we will have $r_{t_1}''/r_{t_1} < M$ for some M . The same inequality holds on this interval for t sufficiently close to t_1 . Applying the (second part of) the Sturm Comparison Theorem to (*) and the equation $y'' + (k^2 - 1)My = 0$, we find that the zeros of $\psi_{k,t}$ must be at least $\varepsilon = \pi / \sqrt{(k^2 - 1)M}$ apart. So if we choose a neighborhood of the graph of ψ_{k,t_1} which intersects the s -axis in intervals of length $< \varepsilon$, then the $\psi_{k,t}$ whose graphs lie in this neighborhood



can have at most n_1 zeros on $(0, 1)$. Thus we have shown that ψ_{k,t_1} actually has exactly n_0 zeros on $(0, 1)$.

But these very same arguments show that for t sufficiently close to t_1 , the function $\psi_{k,t}$ has exactly n_0 zeros on $[0, 1)$ *even for* $t < t_1$. This contradicts the choice of t_1 , and thus shows that indeed $\psi_{k,t_1}(0) = 0$.

Now from the differential equation

$$r_{t_1} \psi_{k,t_1}'' + (k^2 - 1) r_{t_1}'' \psi_{k,t_1} = 0$$

satisfied by ψ_{k,t_1} , and the fact that r_{t_1} is an even function, we easily see that $s \mapsto \psi_{k,t_1}(-s)$ also satisfies the equation. Since $\psi_{k,t_1}(0) = 0$, it follows that $\psi_{k,t_1}(-s) = c \cdot \psi_{k,t_1}(s)$ for some constant c . But this means that the infinitesimal bending determined by ψ_{k,t_1} [on the rotation surface determined by r_{t_1}] is also C^2 at the top pole.

PROBLEMS

1. Let $\alpha, \beta: \mathbb{R} \rightarrow \mathbb{R}^2$ be differentiable functions for which we know the functions

$$\langle \alpha, \beta \rangle, \quad \langle \alpha', \beta \rangle, \quad \langle \alpha, \alpha \rangle, \quad \langle \beta, \beta \rangle,$$

We want to show that we can find α and β once we know $\alpha(0)$ and $\beta(0)$.

(a) We can assume that $\langle \alpha, \alpha \rangle = \langle \beta, \beta \rangle = 1$.

(b) If we write

$$\begin{aligned}\alpha(x) &= (\cos \theta(x), \sin \theta(x)) \\ \beta(y) &= (\cos \phi(y), \sin \phi(y)),\end{aligned}$$

then $\langle \alpha, \beta \rangle$ and $\langle \alpha', \beta \rangle$ determine θ' . Hence $\alpha(0)$ determines α .

(c) Then $\beta(0)$ determines β .

2. (a) Let η be a 1-parameter family of k -forms. Use Proposition 9-10 to show that for every singular $(k+1)$ -cube c we have

$$\int_c (d\eta)^\cdot = \int_{\partial c} \dot{\eta} = \int_c d\dot{\eta}.$$

Conclude that $(d\eta)^\cdot = d\dot{\eta}$.

(b) Give another proof by writing $\eta(u)$ in a coordinate system x^1, \dots, x^n , and noting that $\partial/\partial u$ commutes with $\partial/\partial x^j$.

3. A surface $M \subset \mathbb{R}^3$ is called **isothermal** if it can be covered by isothermal coordinate systems whose parameter lines lie along the lines of curvature.

(a) Surfaces of revolution are isothermal.

(b) Problem 4-8 shows that ellipsoids and hyperboloids of one or two sheets have coordinate systems (u, v) with

$$\langle \ , \ \rangle = (u - v) \left[\frac{u}{f(u)} du \otimes du - \frac{v}{f(v)} dv \otimes dv \right],$$

where the u - and v -parameter lines are lines of curvature. Conclude that these surfaces are isothermal.

(c) Let X_1, X_2 be a moving frame consisting of orthonormal principal vectors. Show that M is isothermal if and only if there is a nowhere zero function α such that

$$d\alpha \wedge \theta^1 + \alpha \omega_1^2 \wedge \theta^2 = 0 = d\alpha \wedge \theta^2 - \alpha \omega_1^2 \wedge \theta^1.$$

Hint: This means that $d(u\theta^1) = d(u\theta^2) = 0$.

4. Let $M \subset \mathbb{R}^3$ be a surface, and $\alpha: [0, 1] \times M \rightarrow \mathbb{R}^3$ a variation as on page 225.

(a) Show that

$$\dot{\psi}_1^3 \wedge \psi_2^3 + \psi_1^3 \wedge \dot{\psi}_2^3 = 0$$

$$d\dot{\psi}_1^3 = \omega_1^2 \wedge \dot{\psi}_2^3$$

$$d\dot{\psi}_2^3 = -\omega_1^2 \wedge \dot{\psi}_1^3.$$

(b) Show that $\dot{H} = 0$ if and only if

$$\dot{\psi}_1^3 \wedge \theta^2 = \dot{\psi}_2^3 \wedge \theta^1.$$

Hint: Use the equation on pg. III.69.

(c) Suppose that the moving frame X_1, X_2 on M consists of principal vectors, so that $\psi_i^3 = k_i \theta^i$, where k_1, k_2 are the principal curvatures. If p is an umbilic, then we always have $\dot{H}(p) = 0$. [*Hint:* Use the first equation of (a).] On the other hand, if $\dot{H} = 0$, and M has no umbilics, then

$$\dot{\psi}_1^3 \wedge \theta^2 = 0 = \dot{\psi}_2^3 \wedge \theta^1.$$

(d) If we write

$$\psi_i^3(t) = \sum_j l_{ij}(t) \theta^j(t), \quad l_{11}(0) = k_1, \quad l_{22}(0) = k_2,$$

then $\dot{H} = 0 \implies \dot{l}_{11} = \dot{l}_{22} = 0$.

(e) Conclude that if $\dot{H} = 0$, then

$$d\dot{l}_{12} \wedge \theta^1 - 2\dot{l}_{12}\omega_2^1 \wedge \theta^1 = 0$$

$$d\dot{l}_{12} \wedge \theta^2 - 2\dot{l}_{12}\omega_1^2 \wedge \theta^1 = 0.$$

Then use **Problem 3** to show that M is isothermal.

5. For vectors v_1, \dots, v_{n-1} and w_1, \dots, w_{n-1} in \mathbb{R}^n , show that

$$\langle v_1 \times \cdots \times v_{n-1}, w_1 \times \cdots \times w_{n-1} \rangle = \det(\langle v_i, w_j \rangle)$$

by noting that both sides are linear in the v_i and w_j .

CHAPTER 13

THE GENERALIZED GAUSS-BONNET THEOREM AND WHAT IT MEANS FOR MANKIND

In previous chapters we have seen that interesting and challenging questions can arise even in the lowest dimensions, and that the methods used to resolve these problems often rely more on ingenuity and hard work than on particularly sophisticated concepts—the proofs may be involved, but they have the satisfying concreteness of geometrical arguments, and something of the charm of antique music. Nevertheless, it is futile to deny the decisive influence which has been wrought upon the shape of modern mathematics by the daemonic spirit of functorial constructions. So it is appropriate that this book end with a topic that represents one of the triumphs of machinery in mathematics. Here, at last, connections in principal bundles play their true predestined role, the invariant form of the Bianchi identities prove their superiority, and connections on arbitrary bundles are frequently invoked. As a final affirmation that we have plunged into the icy stream of modern mathematics, hardly a picture appears.

One of the star attractions of differential geometry is the Gauss-Bonnet Theorem, which for a compact oriented surface M states that

$$\int_M K \, dA = 2\pi \chi(M).$$

Although the curvature K is defined intrinsically in terms of the metric $\langle \cdot, \cdot \rangle$ on M , it can also be defined extrinsically when the metric $\langle \cdot, \cdot \rangle$ on M is induced by an imbedding $M \subset \mathbb{R}^3$. In fact, if $\nu: M \rightarrow S^2$ is the normal map, and da is the volume element of S^2 , then $K \, dA = \nu^*(da)$, so that

$$\int_M K \, dA = \int_M \nu^*(da) = (\deg \nu) \cdot \int_{S^2} da = 4\pi \cdot (\deg \nu).$$

As we indicated in Addendum 2 to Chapter 6, one can prove, without invoking any differential geometry, that $\deg \nu = \frac{1}{2}\chi(M)$, thus proving the Gauss-Bonnet Theorem for the special case where the metric on M comes from an imbedding in \mathbb{R}^3 . Precisely this argument was used by Heinz Hopf [2] in obtaining the

first generalization of the Gauss-Bonnet Theorem. Consider a compact hypersurface $M^n \subset \mathbb{R}^{n+1}$, where n is even. If da denotes the volume element of S^n , then

$$\begin{aligned} \int_M K_n dV &= \int_M \nu^*(da) = (\text{volume of } S^n) \cdot \deg \nu \\ &= \frac{(\text{volume of } S^n)}{2} \cdot \chi(M), \quad \text{by Corollary 6-23.} \end{aligned}$$

Now this result, although proved for a hypersurface of \mathbb{R}^{n+1} , can be formulated for any compact oriented Riemannian n -manifold $(M, \langle \cdot, \cdot \rangle)$ of even dimension n . In fact, we have already noted (pg. IV.69) that

$$(*) \quad K_n = \frac{1}{2^{n/2} n!} \cdot \text{contraction of } \underbrace{(\mathbb{R} \otimes \cdots \otimes \mathbb{R})}_{n/2 \text{ times}} \otimes \mathfrak{E} \otimes \mathfrak{E}.$$

In a coordinate system, we have (pg. IV.69)

$$K_n = \frac{1}{2^{n/2} n!} \sum_{\substack{i_1, \dots, i_n \\ j_1, \dots, j_n}} R_{i_1 i_2 j_1 j_2} \cdots R_{i_{n-1} i_n j_{n-1} j_n} \cdot \frac{\varepsilon^{i_1 \dots i_n}}{\sqrt{\det(g_{ij})}} \cdot \frac{\varepsilon^{j_1 \dots j_n}}{\sqrt{\det(g_{ij})}}.$$

We are thus led to conjecture that we always have

$$\int_M K_n dV = \frac{(\text{volume of } S^n)}{2} \cdot \chi(M),$$

whenever M is a compact oriented Riemannian manifold with n even, where K_n is defined by (*).

For the case where the metric on M comes from an imbedding $M \subset \mathbb{R}^{n+k}$ in some Euclidean space, the result was first proved by Allendoerfer [1] and Fenchel [2]. This was done by considering a closed tubular neighborhood N of M , for which $\partial N \subset \mathbb{R}^{n+k}$ is a hypersurface with a volume element dV , say, and corresponding K_{n+k-1} . We can assume that $n+k-1$ is even (by considering $M \subset \mathbb{R}^{n+k} \subset \mathbb{R}^{n+k+1}$ if necessary), so that the result for hypersurfaces gives

$$\int_{\partial N} K_{n+k-1} dV = \frac{(\text{volume of } S^{n+k-1})}{2} \cdot \chi(\partial N).$$

Now it can be shown without too much difficulty that

$$\chi(\partial N) = \chi(M) \cdot \chi(S^{k-1}) = 2\chi(M).$$

On the other hand, it also works out that

$$\int_{\partial N} \mathbf{K}_{n+k-1} dV$$

can be computed in terms of $\int_M K_n dV$; when the computation is effected, we obtain the correct expression for $\int_M K_n dV$.

At the time of this proof, the Nash imbedding theorem was not yet known. But the Burstin-Janet-Cartan Theorem (Theorem 11-9) was known. In 1943, Allendoerfer and Weil [1] proved a generalization of the Gauss-Bonnet formula for a polyhedral piece of a Riemannian manifold imbedded in Euclidean space; using this, they were able to obtain a proof of the general Gauss-Bonnet Theorem for (C^ω) Riemannian manifolds, by means of a triangulation. Since the Nash imbedding theorem is now available, the earlier result of Allendoerfer and Fenchel implies that the generalized Gauss-Bonnet Theorem holds for all C^∞ manifolds. But a proof of this sort is clearly unsatisfactory, not only because of the difficulty and essentially non-differential geometric nature of Nash's result, but also because an intrinsic theorem ought to have an intrinsic proof. The intrinsic proof was obtained by Chern [2] in 1944. Ensuing developments have led to a much deeper understanding of the fundamentals which are involved here, so that we can now give a completely non-computational proof of this extraordinary theorem. This proof by magic is presented in the first four sections; in the remainder of the Chapter we will contravene the rules of legerdemain, and reveal some of the mechanism behind it.

1. OPERATIONS ON BUNDLES

In the past we have considered numerous structures on vector bundles and principal bundles, but, except in some of the problems for Chapter I.3, we have not yet examined in detail the relationships between different bundles. The simplest relation is that of equivalence \simeq between two vector bundles ξ_1 and ξ_2 over the same base space X . We have also defined the notion of a bundle map from $\xi_1 = \pi_1: E_1 \rightarrow X_1$ to $\xi_2 = \pi_2: E_2 \rightarrow X_2$. This is a pair of continuous maps (\tilde{f}, f) , where $f: X_1 \rightarrow X_2$ and $\tilde{f}: E_1 \rightarrow E_2$; the map \tilde{f} is required to satisfy $\pi_2 \circ \tilde{f} = f \circ \pi_1$, so that \tilde{f} takes fibres of ξ_1 to fibres of ξ_2 , and each map $\tilde{f}: \pi_1^{-1}(x) \rightarrow \pi_2^{-1}(f(x))$ is required to be linear. In this chapter we will redefine the notion of a **bundle map**, by adding the requirement that each $\tilde{f}: \pi_1^{-1}(x) \rightarrow \pi_2^{-1}(f(x))$ be an isomorphism of vector spaces (so ξ_1

and ξ_2 must have the same fibre dimension). Instead of referring to a bundle map (\tilde{f}, f) , we will often say that \tilde{f} is a **bundle map covering** f . If \tilde{f} is a bundle map covering a homeomorphism $f: X_1 \rightarrow X_2$ from X_1 onto X_2 , then we can define $(\tilde{f})^{-1}: E_2 \rightarrow E_1$. Using the local product structure, and the fact that $A \mapsto A^{-1}$ is continuous for $A \in \text{GL}(n, \mathbb{R})$, we easily see that $(\tilde{f})^{-1}$ is continuous, so that $(\tilde{f})^{-1}$ is a bundle map covering f^{-1} . In particular, a bundle map covering the identity map of X is an equivalence.

Consider next two principal bundles $\xi_i = \pi_i: P_i \rightarrow X_i$ ($i = 1, 2$) with the same group G ; we denote the action of G on the right of P_1 and P_2 by the same symbol “ \cdot ”. A **(principal) bundle map** from ξ_1 to ξ_2 is a pair (\tilde{f}, f) , where $f: X_1 \rightarrow X_2$ and $\tilde{f}: P_1 \rightarrow P_2$, such that $\pi_2 \circ \tilde{f} = f \circ \pi_1$, and such that

$$(*) \quad \tilde{f}(u \cdot a) = \tilde{f}(u) \cdot a \quad \text{for all } u \in P \text{ and } a \in G.$$

Notice that condition $(*)$ already implies that \tilde{f} takes fibres to fibres, and thus automatically gives us the map f . We could thus speak simply of a bundle map \tilde{f} . In practice, it is usually more convenient to speak of a **bundle map \tilde{f} covering f** . Note also that $\tilde{f}: \pi_1^{-1}(x) \rightarrow \pi_2^{-1}(f(x))$ is clearly a homeomorphism, since the fibres of P_1 are $\{u \cdot a : a \in G\}$ for fixed u , and similarly for the fibres of P_2 . As before, if \tilde{f} is a bundle map over a homeomorphism $f: X_1 \rightarrow X_2$, then $(\tilde{f})^{-1}$ is a bundle map over f^{-1} . When f is the identity map of X , we call \tilde{f} an **equivalence**, or an **isomorphism**. A principal bundle $\xi = \pi: P \rightarrow X$ is called **trivial** if it is equivalent to the bundle $\pi': X \times G \rightarrow X$, where π' is projection on the first coordinate. As we have already pointed out (pg. II.311), if the principal bundle ξ has a section $s: X \rightarrow P$, then ξ is trivial, for we can define a map $X \times G \rightarrow P$ by $(x, a) \mapsto s(x) \cdot a$.

Recall (pg. II.307) that for every n -dimensional vector bundle $\xi = \pi: E \rightarrow X$ we can define the principal bundle $F(\xi) = \varpi: F(E) \rightarrow X$ of frames of E , with group $\text{GL}(n, \mathbb{R})$, whose fibre $\varpi^{-1}(x)$ is the set of all ordered bases (u_1, \dots, u_n) of the vector space $\pi^{-1}(x)$. If ξ_i are vector bundles over X_i and $\tilde{f}: E_1 \rightarrow E_2$ is a bundle map covering $f: X_1 \rightarrow X_2$, then we clearly have also a principal bundle map $\tilde{f}: F(E_1) \rightarrow F(E_2)$ covering f (this would not be true if we did not require a bundle map to be an isomorphism on each fibre). Conversely, a principal bundle map $\tilde{f}: F(E_1) \rightarrow F(E_2)$ covering f gives rise to a bundle map $\tilde{f}: E_1 \rightarrow E_2$. In fact, given any frame $u = (u_1, \dots, u_n)$ of $\pi_1^{-1}(x)$, there is a unique isomorphism $\pi_1^{-1}(x) \rightarrow \pi_2^{-1}(f(x))$ which takes u_i to the i^{th} member of the frame of $\tilde{f}(u)$. The condition $(*)$ on \tilde{f} insures that this isomorphism is well-defined.

There is one operation which is special for vector bundles. Given two vector bundles $\xi_i = \pi_i: E_i \rightarrow X$ over the same space X , we can form the direct sum $\pi_1^{-1}(x) \oplus \pi_2^{-1}(x)$ [= $\pi_1^{-1}(x) \times \pi_2^{-1}(x)$ as a set] for each $x \in X$. Let

$$E = \bigcup_{x \in X} \pi_1^{-1}(x) \times \pi_2^{-1}(x) \subset E_1 \times E_2,$$

with the topology it has as a subset of $E_1 \times E_2$, and let $\pi: E \rightarrow X$ be the map which takes all elements of $\pi_1^{-1}(x) \times \pi_2^{-1}(x)$ to x (thus $\pi = \pi_i|_E$ for $i = 1, 2$). It is easy to check that $\pi: E \rightarrow X$ is also a vector bundle, whose fibre dimension is the sum of the fibre dimensions of ξ_1 and ξ_2 . This new bundle is called the **Whitney sum** $\xi_1 \oplus \xi_2$ of ξ_1 and ξ_2 . We clearly have

$$\xi_1 \oplus \xi_2 \simeq \xi_2 \oplus \xi_1, \quad (\xi_1 \oplus \xi_2) \oplus \xi_3 \simeq \xi_1 \oplus (\xi_2 \oplus \xi_3).$$

Our next construction works for either a vector bundle or a principal bundle $\xi = \pi: E \rightarrow Y$. Let $f: X \rightarrow Y$ be continuous. We can construct a

$$\begin{array}{ccc} & E & \\ & \downarrow \pi & \\ X & \xrightarrow{f} & Y \end{array}$$

[principal] bundle η over X , and a [principal] bundle map (\tilde{f}, f) from η to ξ , as follows.

$$\begin{array}{ccccc} X \times E \supset E' & \xrightarrow{\tilde{f}} & E & & \\ \downarrow \pi' & & \downarrow \pi & & \\ X & \xrightarrow{f} & Y & & \end{array}$$

Let

$$E' \subset X \times E = \{(x, e) : f(x) = \pi(e)\},$$

and let

$$\pi': E' \rightarrow X \quad \text{be} \quad \pi'((x, e)) = x.$$

Thus the fibre $\pi'^{-1}(x)$ over a point $x \in X$ is just $\{x\} \times \pi^{-1}(f(x))$. In the case of a vector bundle, we use the vector space structure on $\pi^{-1}(f(x))$ to define a vector space structure on $\pi'^{-1}(x)$; in the case of a principal bundle, we use the action of G on $\pi^{-1}(f(x))$ to define the action of G on $\pi'^{-1}(x)$. It is easy to check that $\pi': E' \rightarrow X$ is a vector bundle [principal bundle], and that $\tilde{f}: E' \rightarrow E$ defined by

$$\tilde{f}((x, e)) = e$$

is a [principal] bundle map covering f . The bundle $\pi': E' \rightarrow X$ is denoted by $f^*\xi$, and is called the bundle over X **induced** by f and ξ . If $X \subset Y$ and $i: X \rightarrow Y$ is the inclusion map, then $i^*\xi$ is equivalent to the restriction $\xi|_X$ of ξ to X . If $g: W \rightarrow X$ is another continuous map, then

$$g^*(f^*\xi) \simeq (f \circ g)^*(\xi).$$

Finally, if $\xi = \pi: E \rightarrow Y$ is a vector bundle, then

$$f^*(F(\xi)) \simeq F(f^*\xi).$$

Although we used an explicit construction to define $f^*\xi$, this [principal] bundle can be characterized uniquely, up to equivalence, by the fact that there is a [principal] bundle map covering f from $f^*\xi$ to ξ . Indeed, suppose that $\eta = \pi'': E'' \rightarrow X$ is a [principal] bundle, and $\tilde{f}: E'' \rightarrow E$ is a [principal] bundle map covering f . We define $g: E'' \rightarrow E'$ by

$$g(e'') = (f(\pi''(e'')), \tilde{f}(e''));$$

it is easily seen that g is an equivalence of η and $f^*\xi$.

In the case of two vector bundles $\xi_i = \pi_i: E_i \rightarrow Y$ we have

$$f^*(\xi_1 \oplus \xi_2) \simeq f^*(\xi_1) \oplus f^*(\xi_2).$$

The most reasonable way to prove this is to consider the explicit construction of the total space E of $f^*(\xi_1) \oplus f^*(\xi_2)$, and then define a bundle map \tilde{f} covering f from E to the total space of $\xi_1 \oplus \xi_2$. On the other hand, consider two bundles $\xi_i = \pi_i: E_i \rightarrow X_i$, and let $p_i: X_1 \times X_2 \rightarrow X_i$ be the projections on the factors. Then we can form the bundle

$$\xi_1 \times \xi_2 = p_1^*(\xi_1) \oplus p_2^*(\xi_2)$$

over $X_1 \times X_2$; the fibre over (x_1, x_2) is essentially $\pi_1^{-1}(x_1) \oplus \pi_2^{-1}(x_2)$. When $X_1 = X_2 = X$, we have

$$\xi_1 \oplus \xi_2 \simeq \Delta^*(\xi_1 \times \xi_2),$$

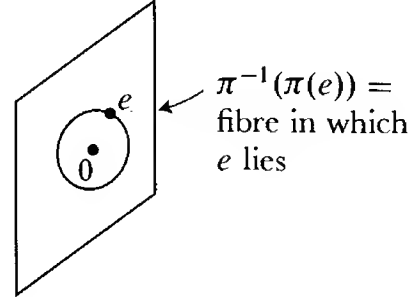
where $\Delta: X \rightarrow X \times X$ is the diagonal map, $\Delta(x) = (x, x)$.

As a somewhat more esoteric example of induced bundles, consider a vector bundle $\xi = \pi: E \rightarrow X$ with a Riemannian metric $\langle \cdot, \cdot \rangle$. Let S be the “sphere bundle”

$$S = \{e \in E : \langle e, e \rangle = 1\},$$

and denote the restriction $\pi|_S$ by $\pi_0: S \rightarrow X$. Then we can form the bundle $\pi_0^*\xi$ over S . We claim that $\pi_0^*\xi$ *always has a nowhere zero section*. To see this, we recall the construction of $\pi_0^*\xi$. For each $e \in S$, the fibre of $\pi_0^*\xi$ over e is just

$$\{e\} \times \pi^{-1}(\pi_0(e)) = \{e\} \times \pi^{-1}(\pi(e)).$$



We define a section s of $\pi_0^*\xi$ by

$$s(e) = (e, e) \in \{e\} \times \pi^{-1}(\pi(e)).$$

Since

$$s(e) \neq (e, 0) = 0 \text{ element of fibre of } \pi_0^*\xi \text{ over } e,$$

this section s is indeed nowhere zero. Similarly, if we regard X as a subspace of E (by considering X as the image of the 0 section), then the induced bundle $(\pi|_{E-X})^*\xi$ over $E-X$ has a nowhere zero section. On the other hand, the bundle $\pi^*\xi$ itself need not have such a section. Indeed, if it does, then the restriction $(\pi^*\xi)|_X$ of $\pi^*\xi$ to X must have a nowhere zero section. But it is clear that $(\pi^*\xi)|_X \simeq \xi$.

The most important result about induced bundles gives a condition under which $f^*\xi \simeq g^*\xi$. As a start towards this result, we note that *any* bundle over $[0, 1]$ is trivial. The proof may be considered as an exercise for the reader; the next Lemma and Theorem establish a more general result.

1. LEMMA. Let ξ be a principal bundle over $X \times [a, b]$. Then every point $x \in X$ has a neighborhood U such that ξ is trivial over $U \times [a, b]$.

PROOF. Each point $(x, t) \in \{x\} \times [a, b]$ has a neighborhood $V \times W$ such that ξ is trivial over $V \times W$. By compactness, finitely many such neighborhoods $V_1 \times W_1, \dots, V_r \times W_r$ cover $\{x\} \times [a, b]$. We claim that the theorem holds with $U = V_1 \cap \dots \cap V_r$. The proof will be by induction on r . For $r = 1$ it is trivial. Assume it holds for $\leq r - 1$ sets. We can clearly choose a point $t_0 \in (a, b)$ such that $[a, t_0]$ and $[t_0, b]$ are each covered by $\leq r - 1$ of the sets $V_i \times W_i$. Then ξ is trivial over sets $U_1 \times [a, t_0]$ and $U_2 \times [t_0, b]$. This means that there is a section s of ξ over $U_1 \times [a, t_0]$ and a section σ of ξ over $U_2 \times [t_0, b]$. On $(U_1 \cap U_2) \times \{t_0\}$ we have

$$s(x, t_0) = \sigma(x, t_0) \cdot a(x)$$

for a continuous function $x \mapsto a(x) \in G$. Then we can define a section \bar{s} on $(U_1 \cap U_2) \times [a, b]$ by

$$\bar{s}(x, t) = \begin{cases} s(x, t) & a \leq t \leq t_0 \\ \sigma(x, t) \cdot a(x) & t_0 \leq t \leq b. \end{cases} \blacklozenge$$

Now let $j: X \times \{1\} \rightarrow X \times [0, 1]$ be the inclusion, and let $p: X \times [0, 1] \rightarrow X \times \{1\}$ be $p(x, t) = (x, 1)$.

2. THEOREM. If $\xi = \pi: P \rightarrow X \times [0, 1]$ is a principal bundle, and X is paracompact, then

$$\xi \simeq p^* j^* \xi \simeq p^*(\xi|X \times \{1\}).$$

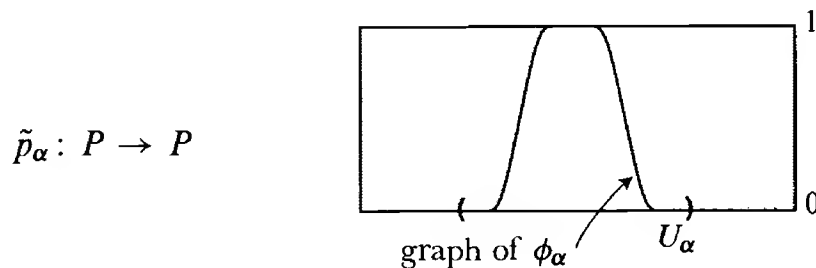
PROOF. We want to show that there is a bundle map $\tilde{p}: P \rightarrow \pi^{-1}(X \times \{1\})$ covering p .

$$\begin{array}{ccc} P & \xrightarrow{\tilde{p}} & \pi^{-1}(X \times \{1\}) \\ \downarrow & & \downarrow \\ X \times [0, 1] & \xrightarrow{p} & X \times \{1\} \end{array}$$

By Lemma 1, there is an open cover $\{U_\alpha\}$ of X such that ξ is trivial on $U_\alpha \times [0, 1]$. We can assume that $\{U_\alpha\}$ is locally finite, by taking a refinement if necessary. By Theorem I.2-15 we can choose a partition of unity $\{\phi_\alpha\}$ with support $\phi_\alpha \subset U_\alpha$ (the Theorem is stated for a manifold, but holds for any normal space X if we only want the functions ϕ_α to be continuous). Let s_α be a section of ξ over $U_\alpha \times [0, 1]$. Consider the map, from $\pi^{-1}(U_\alpha \times [0, 1])$ to itself, defined by

$$s_\alpha(x, t) \mapsto s_\alpha(x, \min(t + \phi_\alpha(x), 1)).$$

This map is the identity on $(\text{boundary } U_\alpha) \times [0, 1]$. So we can extend it continuously to $X \times [0, 1]$ by making it the identity on $(X - U_\alpha) \times [0, 1]$. Thus we obtain a map



which is a bundle map from ξ to the part of ξ over the shaded set in the figure.

Suppose first that there are at most countably many such maps, $\tilde{p}_1, \tilde{p}_2, \dots$. Define

$$\tilde{p} = \tilde{p}_1 \circ \tilde{p}_2 \circ \dots$$

This possibly infinite composition makes sense, since all but finitely many \tilde{p}_i are the identity in a neighborhood of any point. Clearly \tilde{p} is the desired bundle map.

Even if there are uncountably many maps $\{\tilde{p}_\alpha : \alpha \in A\}$, the procedure is the same. Choose any ordering on A (not necessarily a well-ordering) and define \tilde{p} to be the composition of the \tilde{p}_α in the order given by A ; in a neighborhood of any point, only finitely many \tilde{p}_α are not the identity map, so this makes sense. ♦

In practice, it is more convenient to work with a slight restatement, and application, of Theorem 2. Let $i_t : X \rightarrow X \times [0, 1]$ be $i_t(x) = (x, t)$.

3. COROLLARY. If ξ is a principal bundle over $X \times [0, 1]$, and X is paracompact, then

$$i_0^* \xi \simeq i_1^* \xi.$$

PROOF. Let $q : X \times [0, 1] \rightarrow X$ be the projection $q(x, t) = x$. Then $j \circ p = i_1 \circ q$. So, by Theorem 2,

$$(1) \quad \xi \simeq p^* j^* \xi \simeq (j \circ p)^* \xi = (i_1 \circ q)^* \xi \simeq q^* i_1^* \xi.$$

On the other hand, we also have $q \circ i_0 = \text{identity}$. Consequently, equation (1) gives

$$i_0^* \xi \simeq i_0^* q^* i_1^* \xi \simeq [i_1 \circ (q \circ i_0)]^* \xi \simeq i_1^* \xi. \quad \spadesuit$$

From this we immediately obtain the result toward which we have been aiming.

4. THEOREM (THE COVERING HOMOTOPY THEOREM). If η is a principal bundle over Y and $f, g : X \rightarrow Y$ are homotopic, with X paracompact, then $f^* \eta \simeq g^* \eta$. The same result holds if η is a vector bundle.

PROOF. Let $H : X \times [0, 1] \rightarrow Y$ be a map with

$$H \circ i_0 = f \quad \text{and} \quad H \circ i_1 = g.$$

Applying Corollary 3 to $\xi = H^*(\eta)$ over $X \times [0, 1]$, we have

$$\begin{aligned} f^*\eta &= (H \circ i_0)^*\eta \simeq i_0^*(H^*\eta) \\ &\simeq i_1^*(H^*\eta) \simeq (H \circ i_1)^*\eta = g^*\eta. \end{aligned}$$

When η is vector bundle we have, by the remark on page 268,

$$F(f^*\eta) \simeq f^*(F(\eta)) \simeq g^*(F(\eta)) \simeq F(g^*\eta).$$

By the remark on page 266, this implies that $f^*\eta \simeq g^*\eta$. ♦

As a particular case of Theorem 4, note that if X is paracompact and contractible, so that the identity map $1: X \rightarrow X$ is homotopic to a constant map c , then $\xi \simeq 1^*\xi \simeq c^*\xi$, which is trivial. So any principal bundle or vector bundle over X is trivial (compare with remark 3 on pg. I.474).

In applying these results, we will usually be interested only in vector bundles. But in one instance principal bundles will be used. Let $\xi = \pi: E \rightarrow X$ be a vector bundle, and let $\langle \cdot, \cdot \rangle$ be a Riemannian metric on ξ . As in Chapter 7, we can consider the principal bundle $O(\xi) = \varpi: O(E) \rightarrow X$ with group $O(n)$, whose fibre $\varpi^{-1}(x)$ is the set of all frames of $\pi^{-1}(x)$ which are orthonormal with respect to $\langle \cdot, \cdot \rangle$. If $\langle \cdot, \cdot \rangle'$ is another Riemannian metric on ξ , then we have another principal bundle $O'(\xi) = \varpi': O'(E) \rightarrow X$, consisting of frames which are orthonormal with respect to $\langle \cdot, \cdot \rangle'$.

5. COROLLARY. If $\xi = \pi: E \rightarrow X$ is a vector bundle with two Riemannian metrics $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle'$, then $O(\xi) \simeq O'(\xi)$.

PROOF. Let $q: X \times [0, 1] \rightarrow X$ be the projection $q(x, t) = x$, and consider the bundle $q^*\xi$ over $X \times [0, 1]$. The fibre of $q^*\xi$ over (x, t) is

$$\{(x, t)\} \times \pi^{-1}(x).$$

The inner products $\langle \cdot, \cdot \rangle_x$ and $\langle \cdot, \cdot \rangle'_x$ on $\pi^{-1}(x)$ give us an inner product

$$t\langle \cdot, \cdot \rangle_x + (1-t)\langle \cdot, \cdot \rangle'_x$$

on $\pi^{-1}(x)$. Using this inner product on the fibre $\{(x, t)\} \times \pi^{-1}(x)$, we obtain a Riemannian metric $\langle \cdot, \cdot \rangle$ on $q^*\xi$, and we can consider the corresponding principal bundle $O(q^*\xi)$. If $i_t: X \rightarrow X \times [0, 1]$ is $i_t(x) = (x, t)$, then clearly

$$i_0^* O(q^*\xi) \simeq O(\xi) \quad \text{and} \quad i_1^* O(q^*\xi) \simeq O'(\xi).$$

So the result follows from Corollary 3. ♦

2. GRASSMANNIANS AND UNIVERSAL BUNDLES

We have defined projective n -space \mathbb{P}^n to be the set of all pairs $\{p, -p\}$ for $p \in S^n \subset \mathbb{R}^{n+1}$. We could also have defined \mathbb{P}^n to be the set of all lines through 0 in \mathbb{R}^{n+1} , since each such line intersects S^n in a set $\{p, -p\}$. More generally, we define the **Grassmannian manifold** $G_n(\mathbb{R}^N)$ to be the set of all n -dimensional subspaces of \mathbb{R}^N (we will always assume that $N > n$). In order to topologize $G_n(\mathbb{R}^N)$, we consider first the **Stiefel manifold** $V_n(\mathbb{R}^N)$ consisting of all n -tuples

$$(v_1, \dots, v_n) \in \mathbb{R}^N \times \dots \times \mathbb{R}^N$$

for which v_1, \dots, v_n are linearly independent. Clearly $V_n(\mathbb{R}^N)$ is an open subset of $\mathbb{R}^N \times \dots \times \mathbb{R}^N$. We can define a map

$$\rho: V_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^N)$$

by letting

$$\rho((v_1, \dots, v_n)) = \text{subspace of } \mathbb{R}^N \text{ spanned by } v_1, \dots, v_n.$$

We give $G_n(\mathbb{R}^N)$ the quotient topology for this map—thus $\mathcal{U} \subset G_n(\mathbb{R}^N)$ is open if and only if $\rho^{-1}(\mathcal{U}) \subset V_n(\mathbb{R}^N)$ is open.

We can also consider the subspace $V_n^O(\mathbb{R}^N) \subset V_n(\mathbb{R}^N)$ consisting of n -tuples $(v_1, \dots, v_n) \in V_n(\mathbb{R}^N)$ which are orthonormal. If $\rho_0 = \rho|_{V_n^O(\mathbb{R}^N)}$, then the diagram

$$\begin{array}{ccccc} V_n^O(\mathbb{R}^N) & \xrightarrow{i} & V_n(\mathbb{R}^N) & \xrightarrow{g} & V_n^O(\mathbb{R}^N) \\ & \searrow \rho_0 & \downarrow \rho & \swarrow \rho_0 & \\ & & G_n(\mathbb{R}^N) & & \end{array}$$

commutes, where i is the inclusion map, and $g((v_1, \dots, v_n))$ is the n -tuple in $V_n^O(\mathbb{R}^N)$ which results by applying the Gram-Schmidt orthonormalization process to v_1, \dots, v_n . From this diagram it is easy to see that the topology on $G_n(\mathbb{R}^N)$ can also be described as the quotient topology for ρ_0 . Since $V_n^O(\mathbb{R}^N)$ is compact, this shows that $G_n(\mathbb{R}^N)$ is also compact.

There is yet a third description of the topology of $G_n(\mathbb{R}^N)$ which will be important later on. Consider the orthogonal group $O(N)$. If $W_0 \subset \mathbb{R}^N$ is the n -dimensional subspace spanned by e_1, \dots, e_n , then we can define a map

$$O(N) \xrightarrow{\sigma} G_n(\mathbb{R}^N)$$

by

$$\sigma(A) = A(W_0) \in G_n(\mathbb{R}^N).$$

The following diagram then commutes,

$$\begin{array}{ccc} V_n^O(\mathbb{R}^N) & \xrightarrow{\rho} & G_n(\mathbb{R}^N) \\ \alpha \uparrow & \nearrow \sigma & \\ O(N) & & \end{array}$$

where α is the continuous map defined by

$$\alpha(A) = (A(e_1), \dots, A(e_n)) \in V_n^O(\mathbb{R}^N).$$

We thus see that if $\mathcal{U} \subset G_n(\mathbb{R}^N)$ is any set, then

$$\begin{aligned} (1) \quad \rho^{-1}(\mathcal{U}) \text{ is open} &\implies \alpha^{-1}(\rho^{-1}(\mathcal{U})) \text{ is open} \\ &\implies \sigma^{-1}(\mathcal{U}) \text{ is open.} \end{aligned}$$

Notice moreover that we clearly have

$$\alpha(\sigma^{-1}(\mathcal{U})) \subset \rho^{-1}(\mathcal{U}).$$

In addition, the map α is *onto* $V_n^O(\mathbb{R}^N)$, which easily implies that we actually have

$$\alpha(\sigma^{-1}(\mathcal{U})) = \rho^{-1}(\mathcal{U}).$$

Finally, the map α is an *open* map, so if $\mathcal{U} \subset G_n(\mathbb{R}^N)$, then

$$\begin{aligned} (2) \quad \sigma^{-1}(\mathcal{U}) \text{ is open} &\implies \alpha(\sigma^{-1}(\mathcal{U})) \text{ is open} \\ &\implies \rho^{-1}(\mathcal{U}) \text{ is open.} \end{aligned}$$

From (1) and (2) we see that the topology on $G_n(\mathbb{R}^N)$ can also be described as the quotient topology for σ .

This description of $G_n(\mathbb{R}^N)$ is useful for the following reason. The set

$$\sigma^{-1}(W_0) = \{A \in O(N) : A(W_0) = W_0\}$$

is easily seen to consist of all $N \times N$ matrices of the form

$$\begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix} \quad \text{for } C \in O(n) \text{ and } D \in O(N-n).$$

For convenience, this group of matrices is usually denoted by $O(n) \times O(N - n)$. Now any other element of $G_n(\mathbb{R}^N)$ is of the form $B(W_0)$ for some $B \in O(N)$, and

$$\begin{aligned}\sigma^{-1}(B(W_0)) &= \{A \in O(N) : A(W_0) = B(W_0)\} \\ &= \{A \in O(N) : B^{-1}A(W_0) = W_0\} \\ &= \{A \in O(N) : B^{-1}A \in O(n) \times O(N - n)\} \\ &= \text{the left coset } B \cdot (O(n) \times O(N - n)).\end{aligned}$$

Thus we can identify $G_n(\mathbb{R}^N)$ with the left coset space

$$O(N)/O(n) \times O(N - n),$$

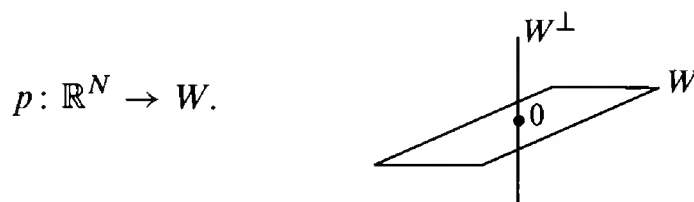
where this quotient space is given the quotient topology for the natural projection map

$$O(N) \longrightarrow O(N)/O(n) \times O(N - n).$$

In section 6 we will study in greater detail the quotient spaces G/H of a Lie group by a closed subgroup, and show that G/H is always a Hausdorff C^∞ manifold of dimension $\dim G - \dim H$. Thus $G_n(\mathbb{R}^N)$ will be a C^∞ manifold of dimension

$$\frac{N(N-1)}{2} - \left\{ \frac{n(n-1)}{2} + \frac{(N-n)(N-n-1)}{2} \right\} = n(N-n).$$

We can also describe this manifold structure on $G_n(\mathbb{R}^N)$ directly as follows. For any $W \in G_n(\mathbb{R}^N)$, consider the orthogonal complement $W^\perp \subset \mathbb{R}^N$. The decomposition $\mathbb{R}^N = W \oplus W^\perp$ determines an orthogonal projection



Let $\mathcal{U} \subset G_n(\mathbb{R}^N)$ be the set of all n -dimensional subspaces V with $V \cap W^\perp = \{0\}$, so that $p: V \rightarrow W$ is an isomorphism. Clearly $\rho^{-1}(\mathcal{U}) \subset V_n(\mathbb{R}^N)$ is open, so \mathcal{U} is an open subset of $G_n(\mathbb{R}^N)$. Now let w_1, \dots, w_n be a fixed orthonormal basis for W , and let w_{n+1}, \dots, w_N be a fixed orthonormal basis for W^\perp . For every $V \in \mathcal{U}$, there are unique $v_1, \dots, v_n \in V$ with $p(v_i) = w_i$, and these v_i can be written uniquely as

$$(*) \quad v_i = w_i + \sum_{j=n+1}^N a_{ij}(V) \cdot w_j.$$

The one-one map

$$V \mapsto (a_{ij}(V))$$

takes \mathcal{U} onto the set of $n \times (N - n)$ matrices. This map is continuous, since the v_i depend continuously on V ; moreover, the inverse map is

$$(a_{ij}) \mapsto \text{space spanned by the } w_i + \sum_{j=n+1}^N a_{ij} \cdot w_j,$$

which is also continuous. Thus we have mapped \mathcal{U} homeomorphically onto $\mathbb{R}^{n(N-n)}$. We leave it to the reader to check that any two such homeomorphisms are C^∞ -related. Thus $G_n(\mathbb{R}^N)$ is a C^∞ manifold. The reader may also check that the map

$$G_n(\mathbb{R}^{m+n}) \rightarrow G_m(\mathbb{R}^{m+n})$$

defined by taking an n -plane W to its orthogonal m -plane W^\perp is a diffeomorphism.

Over the Grassmannian manifold $G_n(\mathbb{R}^N)$ there is a natural n -dimensional bundle $\gamma^n(\mathbb{R}^N)$, constructed as follows. The total space $E(\gamma^n(\mathbb{R}^N))$ of the bundle will be the subset of $G_n(\mathbb{R}^N) \times \mathbb{R}^N$ consisting of all pairs

$$(W, w) \in G_n(\mathbb{R}^N) \times \mathbb{R}^N \quad \text{such that } w \in W,$$

and the projection map $\pi: E(\gamma^n(\mathbb{R}^N)) \rightarrow G_n(\mathbb{R}^N)$ will be $\pi((W, w)) = W$. Thus the fibre $\pi^{-1}(W)$ over the point W of $G_n(\mathbb{R}^N)$ will just be W itself—more precisely, it will be

$$\{(W, w) : w \in W\}.$$

The vector space structure on $\pi^{-1}(W)$ is defined by using the vector space structure on the subspace $W \subset \mathbb{R}^N$; thus

$$\begin{aligned} (W, w_1) + (W, w_2) &= (W, w_1 + w_2) \\ a \cdot (W, w) &= (W, aw). \end{aligned}$$

To show that $\gamma^n(\mathbb{R}^N)$ satisfies the local triviality condition, we consider a point $W \in G_n(\mathbb{R}^N)$, the orthogonal complement W^\perp , the corresponding projection $p: \mathbb{R}^N \rightarrow W$ and the open set $\mathcal{U} \subset G_n(\mathbb{R}^N)$ consisting of all V with $V \cap W^\perp = \{0\}$. Now we can define a map

$$\pi^{-1}(\mathcal{U}) \rightarrow \mathcal{U} \times W \approx \mathcal{U} \times \mathbb{R}^n$$

by taking

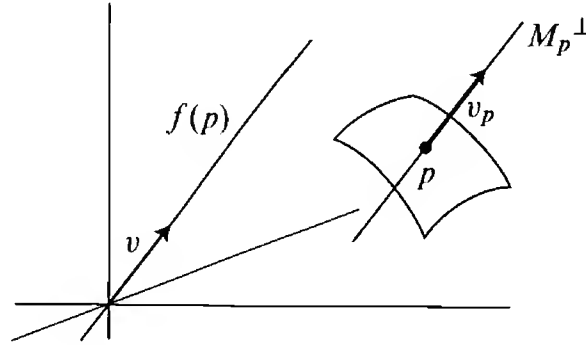
$$(V, v) \mapsto (V, p(v)).$$

This map is easily seen to be a diffeomorphism, and is an isomorphism on each fibre, so $\gamma^n(\mathbb{R}^N)$ is a smooth vector bundle over $G_n(\mathbb{R}^N)$.

Notice that for $M > N$ there is a natural map $\alpha: G_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^M)$, since an n -dimensional subspace of \mathbb{R}^N can be considered as an n -dimensional subspace of \mathbb{R}^M . There is also an obvious map $\tilde{\alpha}: E(\gamma^n(\mathbb{R}^N)) \rightarrow E(\gamma^n(\mathbb{R}^M))$ such that $(\tilde{\alpha}, \alpha)$ is a bundle map from $\gamma^n(\mathbb{R}^N)$ to $\gamma^n(\mathbb{R}^M)$. Thus

$$\gamma^n(\mathbb{R}^N) \simeq \alpha^*(\gamma^n(\mathbb{R}^M)).$$

Now consider a C^∞ manifold M^n immersed in \mathbb{R}^{n+1} . Since M need not be orientable, we may not be able to define the normal map $v: M^n \rightarrow S^n$. But we can certainly define a map $f: M \rightarrow \mathbb{P}^n = G_1(\mathbb{R}^{n+1})$, by taking $p \in M$ to the 1-dimensional subspace of \mathbb{R}^{n+1} which is parallel to the line $M_p^\perp \subset \mathbb{R}^{n+1}_p$.



We can also define a map \tilde{f} from the normal bundle $\text{Nor } M$ of M into the total space of $\gamma^1(\mathbb{R}^{n+1})$ by sending $v_p \in M_p^\perp$ to $(f(p), v)$. Thus we have a bundle map (\tilde{f}, f) from the normal bundle $\text{Nor } M$ to the bundle $\gamma^1(\mathbb{R}^{n+1})$; consequently, the normal bundle $\text{Nor } M$ is equivalent to $f^*(\gamma^1(\mathbb{R}^{n+1}))$. It is even more interesting to look at the map f from M into the diffeomorphic manifold $G_n(\mathbb{R}^{n+1})$ defined by $f(p) =$ subspace of \mathbb{R}^{n+1} parallel to M_p . For then we can define $\tilde{f}: TM \rightarrow E(\gamma^n(\mathbb{R}^{n+1}))$ by sending $v_p \in M_p$ to $(f(p), v)$. Thus we see that the tangent bundle TM is equivalent to $f^*(\gamma^n(\mathbb{R}^{n+1}))$. Moreover, this construction can be generalized. By Proposition I.2-17, any compact n -manifold M can be considered as a submanifold $M^n \subset \mathbb{R}^N$ for some N . Define $f: M \rightarrow G_n(\mathbb{R}^N)$ by $f(p) =$ subspace of \mathbb{R}^N parallel to M_p , and define $\tilde{f}: TM \rightarrow E(\gamma^n(\mathbb{R}^N))$ by sending $v_p \in M_p$ to $(f(p), v)$. Then (\tilde{f}, f) is a bundle map from TM to the bundle $\gamma^n(\mathbb{R}^N)$. Thus the tangent bundle TM is equivalent to $f^*(\gamma^n(\mathbb{R}^N))$. Actually, this holds for *all* bundles.

6. THEOREM. Let $\xi = \pi: E \rightarrow X$ be an n -dimensional bundle over a compact Hausdorff space X . Then for sufficiently large N there is a map $f: X \rightarrow G_n(\mathbb{R}^N)$ such that $\xi \simeq f^*(\gamma^n(\mathbb{R}^N))$.

If X is a smooth manifold and ξ is a smooth bundle, then f can be chosen to be a smooth map.

PROOF. Let U_1, \dots, U_r be open sets covering X such that each $\xi|_{U_i}$ is trivial. The Shrinking Lemma (Theorem I.2-14) holds for the cover U_1, \dots, U_r of X , for the proof merely uses the fact that X is normal. So there is an open cover V_1, \dots, V_r of X with $\bar{V}_i \subset U_i$. Similarly, there is an open cover W_1, \dots, W_r of X with $\bar{W}_i \subset V_i$. Let $\phi_i: X \rightarrow \mathbb{R}$ be a continuous function which is 1 on \bar{W}_i and 0 outside of V_i .

By assumption on the U_i , there are equivalences

$$t_i: \pi^{-1}(U_i) \rightarrow U_i \times \mathbb{R}^n.$$

Composing with the projections $U_i \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, we thus obtain maps

$$\tau_i: \pi^{-1}(U_i) \rightarrow \mathbb{R}^n$$

which are isomorphism on each fibre. Define $\tau_i': E \rightarrow \mathbb{R}^n$ by

$$\tau_i'(e) = \begin{cases} 0 & \pi(e) \notin V_i \\ \phi_i(\pi(e)) \cdot \tau_i(e) & \pi(e) \in U_i. \end{cases}$$

These maps are linear on each fibre, but not one-one on all fibres. Now define

$$T: E \rightarrow \mathbb{R}^n \oplus \dots \oplus \mathbb{R}^n \approx \mathbb{R}^{rn}$$

by

$$T(e) = (\tau_1'(e), \dots, \tau_r'(e)).$$

Then T is linear and one-one on all fibres, so each set $T(\pi^{-1}(x))$ is an n -plane in \mathbb{R}^{rn} . Defining $f: X \rightarrow G_n(\mathbb{R}^{rn})$ and $\tilde{f}: E \rightarrow E(\gamma^n(\mathbb{R}^{rn}))$ by

$$\begin{aligned} f(x) &= T(\pi^{-1}(x)) = \{T(e) : e \in \pi^{-1}(x)\} \in G_n(\mathbb{R}^{rn}) \\ \tilde{f}(e) &= (f(\pi(e)), T(e)) \in E(\gamma^n(\mathbb{R}^{rn})), \end{aligned}$$

it is easily checked that (\tilde{f}, f) is a bundle map from ξ to $\gamma^n(\mathbb{R}^{rn})$.

When ξ is a smooth bundle, we choose the ϕ_i to be smooth, and then f will also be smooth. ♦

The map $f: X \rightarrow G_n(\mathbb{R}^N)$ of Theorem 6 cannot be unique, for Theorem 4 shows that $f^*(\gamma^n(\mathbb{R}^N)) \simeq g^*(\gamma^n(\mathbb{R}^N))$ whenever f and g are homotopic. But in a certain sense this is the only extent to which the representation fails to be unique:

7. THEOREM. Let $f_0, f_1: X \rightarrow G_n(\mathbb{R}^N)$ be two maps such that

$$f_0^* \gamma^n(\mathbb{R}^N) \simeq f_1^* \gamma^n(\mathbb{R}^N).$$

Consider the natural inclusion $\alpha: G_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^M)$, for $M \geq 2N$. Then $\tilde{f}_0 = \alpha \circ f_0$ and $\tilde{f}_1 = \alpha \circ f_1$ are homotopic.

If f_0 and f_1 are smooth maps, then \tilde{f}_0 and \tilde{f}_1 are smoothly homotopic.

PROOF. For each $x \in X$ we have two n -planes $f_0(x)$ and $f_1(x) \in G_n(\mathbb{R}^M)$. By assumption, there is a bundle map from the total space of $f_0^*(\gamma^n(\mathbb{R}^N))$ to the total space of $f_1^*(\gamma^n(\mathbb{R}^N))$. Recalling how these bundles are defined, we see that for each $x \in X$ we have an isomorphism

$$h(x): f_0(x) \rightarrow f_1(x),$$

depending continuously on x . First we consider a

Special Case. For all $x \in X$ and all non-zero $v \in f_0(x)$, the vector $h(x)(v)$ is never a negative multiple of v .

In this case, for each $t \in [0, 1]$ we define a map $h_t(x): f_0(x) \rightarrow \mathbb{R}^N$ by

$$h_t(x) = (1 - t) \cdot \text{identity} + t \cdot h(x).$$

For the image $h_t(x)(f_0(x)) \subset \mathbb{R}^N$ we have

$$h_0(x)(f_0(x)) = f_0(x)$$

$$h_1(x)(f_0(x)) = f_1(x).$$

We define a homotopy f_t between f_0 and f_1 by

$$f_t(x) = h_t(x)(f_0(x)).$$

The assumption in our special case insures that each $h_t(x)$ is one-one on $f_0(x)$, so that we have $f_t(x) \in G_n(\mathbb{R}^N)$. It is not hard to check that $(x, t) \mapsto f_t(x)$ is continuous on $X \times [0, 1]$, and is therefore the desired homotopy.

General Case. In the general case, the above construction does not work. Moreover, it may happen that the hypothesis for the special case will never occur even when we replace f_0 by some homotopic map f_0' . It is necessary to look at the compositions $\tilde{f}_0, \tilde{f}_1: X \rightarrow G_n(\mathbb{R}^M)$. Note that

$$\begin{aligned} \tilde{f}_i^* \gamma^n(\mathbb{R}^M) &= (\alpha \circ f_i)^* \gamma^n(\mathbb{R}^M) \\ &\simeq f_i^* \alpha^* \gamma^n(\mathbb{R}^M) \\ &\simeq f_i^* \gamma^n(\mathbb{R}^N). \end{aligned}$$

So by hypothesis we have

$$\bar{f}_0^* \gamma^n(\mathbb{R}^M) \simeq \bar{f}_1^* \gamma^n(\mathbb{R}^M).$$

Now since $M \geq 2N$, we can define a map $\mathbb{R}^M \rightarrow \mathbb{R}^M$ by

$$(a_1, \dots, a_N, a_{N+1}, \dots, a_{2N}, \dots) \mapsto (a_{N+1}, \dots, a_{2N}, a_1, \dots, a_N, \dots).$$

This induces a map $S: G_n(\mathbb{R}^M) \rightarrow G_n(\mathbb{R}^M)$, which is homotopic to the identity.

Thus $S \circ \bar{f}_1 \simeq \bar{f}_1$, so

$$\bar{f}_0^* \gamma^n(\mathbb{R}^M) \simeq \bar{f}_1^*(\gamma^n(\mathbb{R}^M)) \simeq (S \circ \bar{f}_1)^* \gamma^n(\mathbb{R}^M).$$

But \bar{f}_0 and $S \circ \bar{f}_1$ clearly satisfy the hypotheses of the special case. So we have

$$\bar{f}_0 \simeq S \circ \bar{f}_1 \simeq \bar{f}_1.$$

If f_0 and f_1 are smooth, so that \bar{f}_0 and \bar{f}_1 are smooth, then the homotopy constructed above is also smooth. ♦

In algebraic topology it is customary to consider the union $G_n(\mathbb{R}^\infty)$ of the increasing sequence

$$G_n(\mathbb{R}^{n+1}) \subset G_n(\mathbb{R}^{n+2}) \subset \dots$$

with the “weak topology”: a set $\mathcal{U} \subset G_n(\mathbb{R}^\infty) = \bigcup_l G_n(\mathbb{R}^{n+l})$ is open if and only if $\mathcal{U} \cap G_n(\mathbb{R}^{n+l})$ is open in $G_n(\mathbb{R}^{n+l})$ for all l . There is a natural n -dimensional bundle γ^n over $G_n(\mathbb{R}^\infty)$, defined analogously to $\gamma^n(\mathbb{R}^N)$, and this bundle has the following two properties:

- (A) For every bundle ξ over a paracompact space X there is a map $f: X \rightarrow G_n(\mathbb{R}^\infty)$ such that $\xi \simeq f^*(\gamma^n)$,
- (B) If $f_0, f_1: X \rightarrow G_n(\mathbb{R}^\infty)$ are maps of a paracompact space X into $G_n(\mathbb{R}^\infty)$ with $f_0^* \gamma^n \simeq f_1^* \gamma^n$, then $f_0 \simeq f_1$.

For this reason, γ^n is called the “universal n -dimensional bundle”, and $G_n(\mathbb{R}^\infty)$ is called the “classifying space” for n -dimensional bundles, since equivalence classes of n -dimensional bundles over X are classified by homotopy classes of maps of X into $G_n(\mathbb{R}^\infty)$. Since $G_n(\mathbb{R}^\infty)$ is not a manifold, we do not work with these bundles. Instead we will continue to use the bundles $\gamma^n(\mathbb{R}^N)$, which we also call, somewhat sloppily, “universal bundles”.

All of the preceding discussion can be modified to deal with oriented bundles. Recall that an **orientation** μ for a vector space V is an equivalence class of

ordered bases for V , where $(v_1, \dots, v_n) \sim (w_1, \dots, w_n)$ if and only if the matrix (a_{ij}) defined by $w_i = \sum_j a_{ji} v_j$ has $\det(a_{ij}) > 0$. There are only two such equivalence classes, and the one which is not μ is denoted by $-\mu$. An **oriented** vector space is a pair (V, μ) , where μ is an orientation for V ; the condition $(v_1, \dots, v_n) \in \mu$ is usually expressed by saying that v_1, \dots, v_n is positively oriented (with respect to μ). Given two oriented vector spaces (V, μ) and (W, ν) , we orient $V \oplus W$ by declaring $v_1, \dots, v_n, w_1, \dots, w_m$ to be positively oriented if v_1, \dots, v_n and w_1, \dots, w_m are positively oriented with respect to μ and ν , respectively. Thus the orientation for $W \oplus V$ is $(-1)^{mn}$ times the orientation for $V \oplus W$.

An **orientation** for a bundle $\xi = \pi: E \rightarrow X$ is a collection $\mu = \{\mu_x\}$ of orientations for the fibres $\pi^{-1}(x)$, satisfying an obvious compatibility requirement, while an **oriented bundle** is a pair (ξ, μ) , where μ is an orientation for ξ . An orientation μ for ξ gives us another orientation $-\mu = \{-\mu_x\}$; if X is connected, this is the only other orientation for ξ . Given two oriented bundles (ξ_1, μ_1) and (ξ_2, μ_2) over the same space X , we can define an orientation on the Whitney sum $\xi_1 \oplus \xi_2$ by using the orientation on the direct sums of fibres described in the previous paragraph; thus we can define $(\xi_1 \oplus \xi_2, \mu_1 \oplus \mu_2)$ to be $\xi_1 \oplus \xi_2$ with this orientation. Given an oriented bundle (ξ, μ) over Y , and a continuous map $f: X \rightarrow Y$, there is an obvious way to define an orientation $f^*\mu$ for $f^*\xi$; thus we can define $f^*(\xi, \mu)$ to be the oriented bundle $(f^*\xi, f^*\mu)$.

For two bundles ξ_1, ξ_2 with orientations μ_1, μ_2 , respectively, we can speak of orientation preserving bundle maps and equivalences, or we can simply speak of bundle maps and equivalences between the oriented bundles (ξ_1, μ_1) and (ξ_2, μ_2) . Notice that the oriented bundles (ξ, μ) and $(\xi, -\mu)$ need not be equivalent. For example, if μ is an orientation of the tangent bundle of S^2 , then (TS^2, μ) and $(TS^2, -\mu)$ are not equivalent. In fact, an orientation preserving equivalence from (TS^2, μ) to $(TS^2, -\mu)$ would give us a continuous family of isomorphisms

$$A_p: S^2_p \rightarrow S^2_p, \quad \text{with all } \det A_p < 0.$$

Now a linear transformation $A: V \rightarrow V$ from a 2-dimensional vector space V to itself has two complex eigenvalues λ_1, λ_2 , and if λ_1 is not real, then $\lambda_2 = \overline{\lambda_1}$. But the condition $\det A = \lambda_1 \lambda_2 < 0$ clearly implies that we do not have $\lambda_2 = \overline{\lambda_1}$, so A has two real eigenvalues of opposite signs. Thus we could use the A_p to continuously pick out a 1-dimensional subspace of S^2_p for all $p \in S^2$, by choosing the eigenvectors with the positive eigenvalue for A_p . But such a continuous choice cannot be made, by Problem I.9-7.

We define the **oriented Grassmannian manifold** $\tilde{G}_n(\mathbb{R}^N)$ to be the set of all oriented n -dimensional subspaces of \mathbb{R}^N . We have already defined the map

$\rho: V_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^N)$, where $V_n(\mathbb{R}^N)$ is the Stiefel manifold. We can define a map

$$\tilde{\rho}: V_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^N)$$

by setting

$$\tilde{\rho}((v_1, \dots, v_n)) = (\rho(v_1, \dots, v_n), \mu),$$

where μ is the orientation of $\rho(v_1, \dots, v_n)$ determined by the ordered basis v_1, \dots, v_n . We give $\tilde{G}_n(\mathbb{R}^N)$ the quotient topology for $\tilde{\rho}$. If $\tilde{\rho}_0 = \tilde{\rho}|_{V_n^O(\mathbb{R}^N)}$, then the diagram

$$\begin{array}{ccccc} V_n^O(\mathbb{R}^N) & \xrightarrow{i} & V_n(\mathbb{R}^N) & \xrightarrow{g} & V_n^O(\mathbb{R}^N) \\ & \searrow \tilde{\rho}_0 & \downarrow \tilde{\rho} & \swarrow \tilde{\rho}_0 & \\ & & \tilde{G}_n(\mathbb{R}^N) & & \end{array}$$

commutes; to prove this, one just has to check that the Gram-Schmidt process g preserves orientation. So the topology on $\tilde{G}_n(\mathbb{R}^N)$ could also be described as the quotient topology for $\tilde{\rho}_0$.

Similarly, we can define a map on the special orthogonal group,

$$\tilde{\sigma}: \text{SO}(M) \rightarrow \tilde{G}_n(\mathbb{R}^N),$$

by

$$\tilde{\sigma}(A) = (A(W_0), \mu),$$

where the orientation μ on $A(W_0)$ is that determined by the ordered basis $A(e_1), \dots, A(e_n)$. The diagram

$$\begin{array}{ccc} V_n^O(\mathbb{R}^N) & \xrightarrow{\tilde{\rho}} & \tilde{G}_n(\mathbb{R}^N) \\ \alpha|_{\text{SO}(N)} \uparrow & \nearrow \tilde{\sigma} & \\ \text{SO}(N) & & \end{array}$$

commutes, and $\alpha|_{\text{SO}(N)}$ is onto for $N > n$. So, as before, we see that the topology on $\tilde{G}_n(\mathbb{R}^N)$ can be described as the quotient topology for $\tilde{\sigma}$. It is then easy to see that $\tilde{G}_n(\mathbb{R}^N)$ can be identified with the left coset space

$$\text{SO}(\bar{N}) / \text{SO}(n) \times \text{SO}(N - n).$$

There is a natural map $\tau: \tilde{G}_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^N)$ defined by

$$\tau((W, \mu)) = W.$$

If $W \in G_n(\mathbb{R}^N)$ and $\mathcal{U} \subset G_n(\mathbb{R}^N)$ is the open set on page 275, with w_1, \dots, w_n a fixed orthonormal basis for W , then every $V \in \mathcal{U}$ can be given the orientation $\mu(V)$ determined by the ordered basis v_1, \dots, v_n , where v_i are the unique vectors in V with $p(v_i) = w_i$. The sets

$$\begin{aligned}\mathcal{U}^+ &= \{(V, \mu(V)) : V \in \mathcal{U}\} \\ \mathcal{U}^- &= \{(V, -\mu(V)) : V \in \mathcal{U}\}\end{aligned}$$

are easily seen to be disjoint open subsets of $\tilde{G}_n(\mathbb{R}^N)$. Thus we see that $\tilde{G}_n(\mathbb{R}^N)$ is a smooth manifold, and that $\tau: \tilde{G}_n(\mathbb{R}^N) \rightarrow G_n(\mathbb{R}^N)$ is a 2-fold covering. As a matter of fact, $\tilde{G}_n(\mathbb{R}^N)$ is clearly the oriented 2-fold covering of the non-orientable manifold $G_n(\mathbb{R}^N)$, as described in Problem 8-2.

Over $\tilde{G}_n(\mathbb{R}^N)$ we define an oriented n -dimensional bundle $(\tilde{\gamma}^n(\mathbb{R}^N), \mu)$ as follows. The total space $E(\tilde{\gamma}^n(\mathbb{R}^N))$ consists of all pairs

$$((W, \mu), w) \in \tilde{G}_n(\mathbb{R}^N) \times \mathbb{R}^N \quad \text{such that } w \in W,$$

and we define $\pi((W, \mu), w) = W$. The vector space structure on $\pi^{-1}((W, \mu))$ is defined as before, and we can also define the natural orientation μ on

$$\pi^{-1}((W, \mu)) = \{((W, \mu), w) : w \in W\}$$

by using the orientation μ on W . (We ought to use a symbol like $\mu_{n,N}$, but for simplicity we won't.) Note that $\tau^*\gamma^n(\mathbb{R}^N)$ is equivalent to the bundle $\tilde{\gamma}^n(\mathbb{R}^N)$, where we forget about the orientation μ . For $M > N$ there is a natural map $\alpha: \tilde{G}_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^M)$, with

$$(\tilde{\gamma}^n(\mathbb{R}^N), \mu) \simeq \alpha^*(\tilde{\gamma}^n(\mathbb{R}^M), \mu).$$

8. THEOREM. (1) Let (ξ, μ) be an oriented n -dimensional bundle over a compact Hausdorff space X . Then for sufficiently large N there is a map $f: X \rightarrow \tilde{G}_n(\mathbb{R}^N)$ such that $(\xi, \mu) \simeq f^*(\tilde{\gamma}^n(\mathbb{R}^N), \mu)$. If X is a smooth manifold and ξ is a smooth bundle, then f can be chosen to be a smooth map.

(2) Let $f_0, f_1: X \rightarrow \tilde{G}_n(\mathbb{R}^N)$ be two maps such that $f_0^*(\tilde{\gamma}^n(\mathbb{R}^N), \mu) \simeq f_1^*(\tilde{\gamma}^n(\mathbb{R}^N), \mu)$. Then the compositions $\tilde{f}_0 = \alpha \circ f_0$ and $\tilde{f}_1 = \alpha \circ f_1$ are homotopic, where $\alpha: \tilde{G}_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^M)$ is the natural inclusion, and $M \geq 2N$. If f_0 and f_1 are smooth maps, then \tilde{f}_0 and \tilde{f}_1 are smoothly homotopic.

PROOF. Left to the reader. ♦

3. THE PFAFFIAN

We have already given an intrinsic expression, as well as an expression in terms of a coordinate system, for the function K_n on a compact oriented Riemannian manifold M of even dimension $n = 2m$. But the most important expression for K_n involves the curvature forms Ω_j^i for a positively oriented orthonormal moving frame X_1, \dots, X_n on M . In terms of these forms, we can easily write down the n -form $K_n dV$ which we want to integrate over M . We will be using the symbol $\varepsilon^{j_1 \dots j_n}$ (defined on pg. IV.68); notice that a sum over permutations, like

$$\sum_{\pi \in S_n} A(X_{\pi(1)}, \dots, X_{\pi(n)}),$$

for example, can just as well be written as

$$\sum_{j_1, \dots, j_n} \varepsilon^{j_1 \dots j_n} A(X_{j_1}, \dots, X_{j_n}).$$

Now consider the m -fold wedge product

$$\Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}}.$$

From the definition of \wedge we have (remembering that the Ω_j^i are 2-forms)

$$\begin{aligned} & \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}}(X_1, \dots, X_n) \\ &= \frac{(2 + \dots + 2)!}{2! \dots 2!} \cdot \frac{1}{n!} \sum_{j_1, \dots, j_n} \varepsilon^{j_1 \dots j_n} \Omega_{i_2}^{i_1}(X_{j_1}, X_{j_2}) \dots \Omega_{i_n}^{i_{n-1}}(X_{j_{n-1}}, X_{j_n}) \\ &= \frac{1}{2^{n/2}} \cdot \sum_{j_1, \dots, j_n} \varepsilon^{j_1 \dots j_n} \langle R(X_{j_1}, X_{j_2})X_{i_2}, X_{i_1} \rangle \dots \langle R(X_{j_{n-1}}, X_{j_n})X_{i_n}, X_{i_{n-1}} \rangle \\ &= \frac{1}{2^{n/2}} \cdot \sum_{j_1, \dots, j_n} \varepsilon^{j_1 \dots j_n} R_{i_1 i_2 j_1 j_2} \dots R_{i_{n-1} i_n j_{n-1} j_n} \quad (\text{see pg. II.190}). \end{aligned}$$

So the formula on page 264 gives

$$K_n = \frac{1}{2^{n/2} n!} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \cdot 2^{n/2} \cdot \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}}(X_1, \dots, X_n),$$

and thus

$$K_n dV = \frac{1}{n!} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}}.$$

This computation shows, in particular, that the form on the right does not depend on the choice of the positively oriented orthonormal moving frame X_1, \dots, X_n . It is also possible to prove this fact directly, by the following algebraic considerations.

For an $n \times n$ matrix $A = (a_{ij})$ with $n = 2m$ even, we define the **Pfaffian** $\text{Pf}(A)$ of A by

$$\text{Pf}(A) = \frac{1}{2^m m!} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} a_{i_1 i_2} \cdots a_{i_{n-1} i_n}.$$

It will soon become clear why the factor $1/2^m m!$ should appear. At the moment we can account for the $1/m!$ by observing that our expression has a lot of redundancy in it. Note first that $\varepsilon^{i_1 \dots i_n}$ does not change when we interchange i_{2l-1} and i_{2l} and also i_{2l-1} and i_{2k} ; more generally, it does not change when we perform any permutation of the *pairs* (i_{2l-1}, i_{2l}) . So for any set $P = \{(h_1, k_1), \dots, (h_m, k_m)\}$ of pairs of integers between 1 and n , it makes sense to define

$$\varepsilon(P) = \varepsilon^{h_1 k_1 \dots h_m k_m};$$

it is not necessary to specify any ordering on the pairs (h_i, k_i) in P . Notice also that a permutation of the pairs (i_{2l-1}, i_{2l}) does not change the factor

$$a_{i_1 i_2} \cdots a_{i_{n-1} i_n}.$$

So for each P as above we can define

$$a_P = a_{h_1 k_1} \cdots a_{h_m k_m}.$$

If \mathcal{P} is the collection of all such P , we then clearly have

$$\text{Pf}(A) = \frac{1}{2^m} \sum_{P \in \mathcal{P}} \varepsilon(P) a_P.$$

9. PROPOSITION. Let $n = 2m$ be even. Then for all $n \times n$ matrices A and B we have

$$\text{Pf}(B^t A B) = (\det B) \cdot \text{Pf}(A),$$

where t denotes the transpose. In particular, if $B \in \text{SO}(n)$, then

$$\text{Pf}(B^{-1} A B) = \text{Pf}(A).$$

PROOF. We have

$$\begin{aligned}
 2^m m! \operatorname{Pf}(B^t A B) &= \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \sum_{j_1, \dots, j_n} (b_{j_1 i_1} a_{j_1 j_2} b_{j_2 i_2}) \cdots (b_{j_{n-1} i_{n-1}} a_{j_{n-1} j_n} b_{j_n i_n}) \\
 &= \sum_{j_1, \dots, j_n} \left[\sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} b_{j_1 i_1} \cdots b_{j_n i_n} \right] \cdot a_{j_1 j_2} \cdots a_{j_{n-1} j_n} \\
 &= \sum_{j_1, \dots, j_n} (\varepsilon^{j_1 \dots j_n} \det B) \cdot a_{j_1 j_2} \cdots a_{j_{n-1} j_n} \\
 &= 2^m m! (\det B) \operatorname{Pf}(A). \quad \blacklozenge
 \end{aligned}$$

Proposition 9 was stated for matrices of real numbers, but $\operatorname{Pf}(A)$ can be defined so long as the entries of A are in some commutative algebra \mathcal{A} over \mathbb{R} . It is easy to see that Proposition 9 still holds when A and B have entries in \mathcal{A} ; in fact, the proof works without change. We could also deduce this extended version from the original Proposition by the “principle of extension of algebraic identities”: First consider the ring $\mathbb{R}[A_{ij}, B_{ij}]$ obtained by adjoining commuting indeterminates A_{ij} and B_{ij} to \mathbb{R} . Then the polynomials

$$\operatorname{Pf}(B^t A B) \quad \text{and} \quad (\det B) \cdot \operatorname{Pf}(A)$$

are elements of $\mathbb{R}[A_{ij}, B_{ij}]$. Proposition 9 tells us that these polynomials have the same values on all $(a_{ij}), (b_{ij})$ for $a_{ij}, b_{ij} \in \mathbb{R}$. Therefore they are equal *as polynomials* in the indeterminates A_{ij}, B_{ij} . So these polynomials are equal when we substitute elements a_{ij}, b_{ij} of the algebra \mathcal{A} for the indeterminates A_{ij}, B_{ij} . Q.E.D.

Now let us consider once again a positively oriented orthonormal moving frame $\mathbf{X} = X_1, \dots, X_n$ on M , with curvature forms Ω_j^i . For each $p \in M$, the direct sum

$$\mathcal{A} = \mathbb{R} \oplus \Omega^2(M_p) \oplus \Omega^4(M_p) \oplus \cdots$$

is a commutative algebra over \mathbb{R} , under \wedge . Consequently, it makes sense to write $\operatorname{Pf}(\Omega(p))$, where $\Omega(p)$ is the $n \times n$ matrix $(\Omega_j^i(p))$ of connection 2-forms at p . In fact, we clearly have

$$\operatorname{Pf}(\Omega(p)) = \frac{1}{2^m m!} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \cdots \wedge \Omega_{i_n}^{i_{n-1}}(p).$$

Now if $\mathbf{X}' = \mathbf{X} \cdot a$ is another positively oriented orthonormal moving frame, then $a(p) \in O(n)$, and by Proposition II.7-15 the corresponding curvature forms (Ω'^i_j) satisfy

$$\Omega' = a^{-1} \Omega a.$$

Then Proposition 9, in its extended form, shows that

$$\text{Pf}(\Omega'(p)) = \text{Pf}(a^{-1}(p)\Omega(p)a(p)) = \text{Pf}(\Omega(p));$$

thus the form

$$\sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}}$$

is indeed well-defined.

Later on we will need to know the important algebraic properties of $\text{Pf}(A)$ that hold when A is *skew-symmetric*. In this case, even the expression

$$\text{Pf}(A) = \frac{1}{2^m} \sum_{P \in \mathcal{P}} \varepsilon(P) a_P$$

is redundant, for the term

$$\varepsilon^{i_1 \dots i_n} a_{i_1 i_2} \dots a_{i_{n-1} i_n}$$

is unchanged when we interchange i_{2l-1} and i_{2l} . Let $\mathcal{P}' \subset \mathcal{P}$ be the collection of all $P = \{(h_1, k_1), \dots, (h_m, k_m)\} \in \mathcal{P}$ with $h_i < k_i$ for all i . Then for skew-symmetric A we clearly have

$$\text{Pf}(A) = \sum_{P \in \mathcal{P}'} \varepsilon(P) a_P,$$

which is a polynomial with integer coefficients. (It follows that $\text{Pf}(A)$ can be defined for a skew-symmetric matrix A with entries in any commutative ring \mathcal{A} with unit.)

There is an important canonical form for skew-symmetric matrices, which is merely a reformulation of the following result which has already appeared in Problem I.7-8.

10. PROPOSITION. Let V be an n -dimensional vector space, and let $\alpha \in \Omega^2(V)$. Then there is a basis ϕ_1, \dots, ϕ_n of V^* such that

$$\alpha = (\phi_1 \wedge \phi_2) + \dots + (\phi_{2r-1} \wedge \phi_{2r})$$

for some r . (For $\alpha = 0$ we must allow the vacuous sum, with $r = 0$.) If $\{\phi_i\}$ is the dual basis to $\{v_i\}$, this means that

$$(1) \quad \begin{cases} \alpha(v_{2i-1}, v_{2i}) = -\alpha(v_{2i}, v_{2i-1}) = 1 & \text{for } i \leq r \\ \alpha(v_i, v_j) = 0 & \text{for all other pairs } i, j. \end{cases}$$

PROOF. We use induction on n , the result being trivial for $n = 1$. Assume the result is true for dimensions $< n$, and consider a non-zero $\alpha \in \Omega^2(V)$, where V has dimension n . There exist $v_1, v_2 \in V$ with $\alpha(v_1, v_2) = 1$; let $[v_1, v_2]$ be the subspace generated by v_1 and v_2 . Now consider the subspace $W \subset V$ of all $v \in V$ such that $\alpha(v_1, v) = \alpha(v_2, v) = 0$. This subspace W is $\ker f_1 \cap \ker f_2$ where $f_i: V \rightarrow \mathbb{R}$ is defined by $f_i(v) = \alpha(v_i, v)$. So $\dim W \geq n - 2$. Moreover, we clearly have $W \cap [v_1, v_2] = \{0\}$, so $\dim W = n - 2$ and $V = [v_1, v_2] \oplus W$. Since the result is assumed true for $\alpha|_{W \times W}$, there is a basis v_3, \dots, v_n of W such that (I) holds for these v_i . Then $v_1, v_2, v_3, \dots, v_n$ is the desired basis. ♦

11. COROLLARY. Let $A = (a_{ij})$ be an $n \times n$ skew-symmetric matrix. Then there is a non-singular $n \times n$ matrix B such that

$$B^t A B = \begin{pmatrix} S & & & & \\ & S & & & \\ & & \ddots & & \\ & & & S & \\ & & & & 0 \\ & & & & 0 & \ddots \\ & & & & & & 0 \end{pmatrix},$$

where S is the matrix

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

PROOF. Let e_1, \dots, e_n be the standard basis of \mathbb{R}^n , and define $\alpha \in \Omega^2(\mathbb{R}^n)$ by

$$\alpha(e_i, e_j) = a_{ij}.$$

Let v_1, \dots, v_n be the basis of \mathbb{R}^n given by Proposition 10, and let $B = (b_{ij})$ be the matrix defined by

$$v_i = \sum_{k=1}^n b_{ki} e_k.$$

Then

$$\begin{aligned} \alpha(v_i, v_j) &= \alpha\left(\sum_k b_{ki} e_k, \sum_l b_{lj} e_l\right) \\ &= \sum_{k,l} b_{ki} b_{lj} \alpha(e_k, e_l) \\ &= \sum_{k,l} b_{ki} b_{lj} a_{kl} = (B^t A B)_{ij}, \end{aligned}$$

which gives the desired result. ♦

Using the expression

$$\text{Pf}(A) = \sum_{P \in \mathcal{P}'} \varepsilon(P) a_P$$

for skew-symmetric A , it is easy to compute that

$$\text{Pf} \begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix} = 1.$$

On the other hand, this matrix also has determinant = 1. This gives us

12. COROLLARY. For every skew-symmetric $n \times n$ matrix A with $n = 2m$ even we have

$$\{\text{Pf}(A)\}^2 = \det A.$$

PROOF. It suffices to prove this when $\det A \neq 0$, since both sides are continuous functions of the entries of A and the matrices with non-zero determinant are dense. By Corollary 11, there is a non-singular $n \times n$ matrix B with

$$B^t A B = \begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix}.$$

Then

$$1 = \det \begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix} = (\det B)^2 \cdot \det A,$$

while Proposition 9 gives

$$1 = \text{Pf} \begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix} = (\det B) \text{Pf}(A). \quad \blacklozenge$$

For a skew-symmetric $n \times n$ matrix A with n odd we define

$$\text{Pf}(A) = 0.$$

Note that in this case we have

$$\det A = \det A^t = \det(-A) = (-1)^n \det A \implies \det A = 0.$$

So Corollary 12 still holds.

For even n the skew-symmetry of A is likewise crucial in Corollary 12. If we consider $\det(A_{ij})$ as a polynomial in n^2 independent commuting variables A_{ij} , then $\det(A_{ij})$ is *not* the square of another polynomial. But if we consider $\det(A_{ij})$ as a polynomial in the $n(n-1)/2$ independent commuting variables A_{ij} for $i < j$ [and define $A_{ii} = 0$, $A_{ij} = -A_{ji}$ for $i > j$], then $\det(A_{ij}) = \{\text{Pf}(A)\}^2$ as polynomials in these A_{ij} , since the two polynomials give the same results when applied to all real numbers a_{ij} , $i < j$. Since $\mathbb{R}[A_{ij}]$ is a unique factorization domain, $\text{Pf}(A)$ and $-\text{Pf}(A)$ are the only two polynomials with this property. The principal of extension of algebraic identities shows that $\det A = \{\text{Pf}(A)\}^2$ when A is a $2m \times 2m$ skew-symmetric matrix with entries in any commutative ring \mathcal{A} with unit.

[By working in the ring $\mathbb{R}[A_{ij}]$, we could have produced the Pfaffian in a neat, mysterious, way that avoids all computations. For there is a matrix X with entries in the quotient field of $\mathbb{R}[A_{ij}]$ such that

$$X^t A X = \begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix}.$$

Hence the polynomial $\det A$ in $\mathbb{R}[A_{ij}]$ is the square $(\det X)^{-2}$ in the quotient field of $\mathbb{R}[A_{ij}]$. Since $\mathbb{R}[A_{ij}]$ is a unique factorization domain, this implies that $\det A$ is already a square in $\mathbb{R}[A_{ij}]$. There are only two possible elements $\text{Pf}(A)$ for $\det A$ to be the square of, and we determine $\text{Pf}(A)$ by requiring that Pf have the value $+1$ on

$$\begin{pmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{pmatrix}.$$

Similarly, if we consider the ring $\mathbb{R}[A_{ij}, B_{11}, \dots, B_{nn}]$ in the indeterminants A_{ij} for $i < j$, and B_{ij} for all i, j , then the identity $\det B^t A B = (\det B)^2 \det A$ implies that

$$\text{Pf}(B^t A B) = \pm(\det B) \text{Pf} A,$$

and by choosing $B = I$ we see that the sign must be $+1$.]

As a rather trivial example of the use of polynomial rings to avoid some computations, we prove one more simple, but important property of Pf . If A and B are two square matrices, we will use $A \oplus B$ for the matrix

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

13. COROLLARY. For all skew-symmetric matrices A and B we have

$$\text{Pf}(A \oplus B) = \text{Pf}(A) \cdot \text{Pf}(B).$$

Note: This result holds even when A or B , or both, is of odd order. In that case it says that $\text{Pf}(A \oplus B) = 0$.

PROOF. By Corollary 12 (which holds even for matrices of odd order) we have

$$(1) \quad \{\text{Pf}(A \oplus B)\}^2 = \det(A \oplus B) = \det A \cdot \det B = \{\text{Pf}(A) \text{Pf}(B)\}^2,$$

and thus

$$(2) \quad \text{Pf}(A \oplus B) = \pm \text{Pf}(A) \text{Pf}(B).$$

If A or B is of odd order, then we already have $\text{Pf}(A \oplus B) = 0$. For A and B of even order, we just need to determine the sign in (2). Now the same sign must hold in (2) for all A and B , for we may consider equation (1) as an equation in the ring $\mathbb{R}[A_{ij}, B_{ij}]$ with commuting variables A_{ij}, B_{ij} ($i < j$ in both cases). Letting A and B be of the form $S \oplus \cdots \oplus S$, we see that the $+$ sign always holds. ♦

4. DEFINING THE EULER CLASS IN TERMS OF A CONNECTION

Consider a smooth oriented n -dimensional vector bundle $\xi = \pi: E \rightarrow M$, over a smooth manifold M (of any dimension). For compact orientable M we defined the Euler class $\chi(\xi) \in H^n(M)$ in Chapter I.11. To do this, we first defined the Thom class $U(\xi) \in H_c^n(E)$, and we proved (Theorem I.11-26) that $U(\xi)$ is the unique class whose restriction to each fibre $\pi^{-1}(p)$ is the generator $\nu_p \in H_c^n(\pi^{-1}(p))$ determined by the orientation. From this result we can immediately conclude

14. LEMMA. Let $\xi = \pi: E \rightarrow M$ be a smooth oriented vector bundle over a compact oriented manifold M , and let $f: M' \rightarrow M$ be a smooth map, where M' is also a compact oriented manifold. If E' is the total space of $f^*\xi$, and $\tilde{f}: E' \rightarrow E$ is a bundle map covering f , then

$$\tilde{f}^*(U(\xi)) = U(f^*\xi) \in H_c^n(E').$$

PROOF. Note first that \tilde{f} is proper (the inverse image of a compact set is compact), so \tilde{f}^* does take $H_c^n(E)$ to $H_c^n(E')$. Let $f^*\xi$ be $\pi': E' \rightarrow M'$. If $p' \in M'$ is any point, and $j_{p'}: \pi'^{-1}(p') \rightarrow E'$ is the inclusion, then

$$j_{p'}^* \tilde{f}^* U(\xi) = (\tilde{f} \circ j_{p'})^* U(\xi).$$

If we recall how $f^*\xi$ is defined, we see that $(\tilde{f} \circ j_{p'})^* U(\xi)$ must be the generator of $H_c^n(\pi'^{-1}(p'))$, since $j_{f(p')}^* U(\xi)$ is the generator of $H_c^n(\pi^{-1}(f(p')))$. This shows that $\tilde{f}^* U(\xi)$ must be $U(f^*\xi)$. ♦

We defined the Euler class $\chi(\xi)$ to be $s^* U(\xi)$, for any section s of ξ . In particular, we can choose $s = 0 =$ the zero section. Hence

15. PROPOSITION. Let $\xi = \pi: E \rightarrow M$ be a smooth oriented vector bundle over a compact oriented manifold M , and let $f: M' \rightarrow M$ be a smooth map, where M' is also a compact oriented manifold. Then

$$f^* \chi(\xi) = \chi(f^* \xi) \in H^n(M').$$

PROOF. If $0'$ denotes the zero section of $f^* \xi$, then $\tilde{f} \circ 0' = 0 \circ f$. So

$$\begin{aligned} \chi(f^* \xi) &= (0')^* U(f^* \xi) \\ &= (0')^* \tilde{f}^* U(\xi) \quad \text{by Lemma 14} \\ &= (\tilde{f} \circ 0')^* U(\xi) = (0 \circ f)^* U(\xi) \\ &= f^* 0^* U(\xi) = f^* \chi(\xi). \quad \spadesuit \end{aligned}$$

As a particular consequence, we note a result which we will need later on.

16. COROLLARY. If n is even, then

$$\chi(\tilde{\gamma}^n(\mathbb{R}^N)) \neq 0$$

for all $N > n$.

PROOF. Since $S^n \subset \mathbb{R}^N$ for $N > n$, we have a bundle map $(\tilde{f}, f): TS^n \rightarrow E(\tilde{\gamma}^n(\mathbb{R}^N))$, as on page 277. So

$$\chi(TS^n) = f^* \chi(\tilde{\gamma}^n(\mathbb{R}^N)).$$

But Theorem I.11-30 says that $\chi(TS^n)$ is $\chi(S^n)$ times the fundamental class of S^n , and $\chi(S^n) = 2 \neq 0$. ♦

The Euler class has one further important property which it is not really essential to prove at this point, for it could eventually be derived from other results of this section. Nevertheless, it will motivate much of the argument to come.

17. THEOREM. Let $\xi_i = \pi_i: E_i \rightarrow M$ for $i = 1, 2$ be smooth oriented vector bundles over a compact oriented manifold M . Then $\chi(\xi_1 \oplus \xi_2)$ is the cup product.

$$\chi(\xi_1 \oplus \xi_2) = \chi(\xi_1) \cup \chi(\xi_2).$$

PROOF. The Whitney sum $\xi_1 \oplus \xi_2$ is $\pi: E \rightarrow M$ where $E \subset E_1 \times E_2$ is $\{(e_1, e_2) : \pi_1(e_1) = \pi_2(e_2)\}$. Let $\rho_i: E \rightarrow E_i$ be the restriction of the projection maps $E_1 \times E_2 \rightarrow E_i$. For any $p \in M$, let

$$j: \pi^{-1}(p) \rightarrow E, \quad j_i: \pi_i^{-1}(p) \rightarrow E_i$$

be the inclusions, and let

$$\sigma_i: \pi_1^{-1}(p) \times \pi_2^{-1}(p) \rightarrow \pi_i^{-1}(p)$$

be the projections. Then

$$j_i \circ \sigma_i = \rho_i \circ j.$$

So

$$\begin{aligned} j^*(\rho_1^*U_1(\xi_1) \cup \rho_2^*U_2(\xi_2)) &= (\rho_1 \circ j)^*U_1(\xi) \cup (\rho_2 \circ j)^*U_2(\xi) \\ &= \sigma_1^*j_1^*U_1(\xi) \cup \sigma_2^*j_2^*U_2(\xi) \\ &= \sigma_1^*v_1 \cup \sigma_2^*v_2, \end{aligned}$$

where v_i is the generator of $H_c^{n_i}(\pi_i^{-1}(p))$ determined by the orientation on $\pi_i^{-1}(p)$. But $\sigma_1^*v_1 \cup \sigma_2^*v_2$ is easily seen to be the generator of the group $H_c^{n_1+n_2}(\pi_1^{-1}(p) \times \pi_2^{-1}(p))$ determined by the orientation on $\pi_1^{-1}(p) \times \pi_2^{-1}(p)$ (given on page 281). It follows that

$$\rho_1^*U_1(\xi_1) \cup \rho_2^*U_2(\xi_2) = U(\xi), \quad \text{the Thom class of } \xi.$$

So if s_i are sections of ξ_i , then for the obvious section $s_1 + s_2$ of ξ we have

$$\begin{aligned} \chi(\xi) &= (s_1 + s_2)^*U(\xi) = (s_1 + s_2)^*(\rho_1^*U_1(\xi_1) \cup \rho_2^*U_2(\xi_2)) \\ &= [\rho_1 \circ (s_1 + s_2)]^*U_1(\xi_1) \cup [\rho_2 \circ (s_1 + s_2)]^*U_2(\xi_2) \\ &= s_1^*U_1(\xi_1) \cup s_2^*U_2(\xi_2) \\ &= \chi_1(\xi_1) \cup \chi_2(\xi_2). \quad \blacklozenge \end{aligned}$$

Now we are going to look at principal bundles associated with a smooth oriented n -dimensional vector bundle $\xi = \pi: E \rightarrow M$ over a smooth manifold M . We have already considered the principal bundle $F(\xi)$ of frames of E . If we have a Riemannian metric $\langle \cdot, \cdot \rangle$ for ξ , then, as in Chapter 7, we can consider the bundle $O(E)$ of orthonormal frames, which is a principal bundle with group $O(n)$. Since we will be considering only paracompact manifolds M , we know (see pg. II.342) that there is an Ehresmann connection ω on the bundle $O(E)$. Thus ω is a matrix of 1-forms (ω_j^i) on $O(E)$ taking values in $\mathfrak{o}(n)$; the curvature form $\Omega = D\omega$ is a matrix of 2-forms (Ω_j^i) , also with values in $\mathfrak{o}(n)$. [Since we will seldom be working with TM any more, and never with moving frames, we will not resort to any special symbolism to distinguish the forms ω_j^i, Ω_j^i defined on a bundle from those defined for some moving frame.] As we pointed out in Chapter 7, a connection ω on $O(E)$ is equivalent to a covariant differentiation operator on E which is compatible with the metric $\langle \cdot, \cdot \rangle$. In the case of a general bundle ξ over M there will be many connections compatible with the metric $\langle \cdot, \cdot \rangle$; we cannot single one out by asking for a symmetric connection, as this concept makes sense only for the tangent bundle. Since our bundle ξ is oriented, we can also consider the bundle $SO(E)$ of positively oriented orthonormal frames; if X is connected, it is simply one of the two components of $O(E)$. The group of this bundle is $SO(n)$, whose Lie algebra is also $\mathfrak{o}(n)$. So a connection ω on $SO(n)$ again has values in $\mathfrak{o}(n)$, as does the matrix of 2-forms Ω .

Now let us specialize to the case of a smooth oriented n -dimensional vector bundle $\xi = \pi: E \rightarrow M$ over M , where $n = 2m$ is *even*. If $\langle \cdot, \cdot \rangle$ is a Riemannian metric for ξ , and ω is a connection on the corresponding principal bundle $\varpi: SO(E) \rightarrow M$, then we can consider the n -form

$$2m \cdot m! \text{Pf}(\Omega) = \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}},$$

which is defined *on the bundle* $SO(E)$. The following proof is merely an invariant formulation of an argument presented in the last section.

18. PROPOSITION. There is a unique n -form Λ on M such that

$$\varpi^*(\Lambda) = \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}} = 2^m m! \text{Pf}(\Omega).$$

PROOF. Given $X_1, \dots, X_n \in M_p$, choose some $u \in \varpi^{-1}(p)$, and let $Y_1, \dots, Y_n \in SO(E)_u$ be tangent vectors with $\varpi_* Y_i = X_i$. Clearly Λ must satisfy

$$\Lambda(X_1, \dots, X_n) = 2^m m! \text{Pf}(\Omega)(Y_1, \dots, Y_n),$$

which proves uniqueness. Existence will be demonstrated once we prove that this Λ is well-defined.

Consider first what happens when we take different tangent vectors $Z_1, \dots, Z_n \in \text{SO}(E)_u$ with $\varpi_* Z_i = X_i$. Since $\varpi_*(Y_i - Z_i) = 0$, all $Y_i - Z_i$ are vertical. But $\Omega(Y, Z) = 0$ if either Y or Z is vertical. So we clearly have

$$\begin{aligned} \text{Pf}(\Omega)(Y_1, \dots, Y_n) &= \text{Pf}(\Omega)(Z_1, Y_2, \dots, Y_n) \\ &= \text{Pf}(\Omega)(Z_1, Z_2, Y_3, \dots, Y_n) = \dots \\ &= \text{Pf}(\Omega)(Z_1, \dots, Z_n). \end{aligned}$$

Thus the definition of Λ does not depend on the choice of the Y_i .

Now suppose we choose a different $\bar{u} \in \varpi^{-1}(p)$. Then $\bar{u} = R_A(u) = u \cdot A$ for some $A \in \text{SO}(n)$, and we can let the $\bar{Y}_i \in \text{SO}(E)_{\bar{u}}$ be $\bar{Y}_i = R_{A*} Y_i$. Then

$$\begin{aligned} \text{Pf}(\Omega)(\bar{Y}_1, \dots, \bar{Y}_n) &= \text{Pf}(\Omega)(R_{A*} Y_1, \dots, R_{A*} Y_n) \\ &= \text{Pf}(R_A^* \Omega)(Y_1, \dots, Y_n) \\ &= \text{Pf}(A^{-1} \Omega A)(Y_1, \dots, Y_n) && \text{by Proposition II.8-11} \\ &= \text{Pf}(\Omega)(Y_1, \dots, Y_n) && \text{by Proposition 9. } \spadesuit \end{aligned}$$

We also have the following result, which is automatic when ξ is the tangent bundle of M .

19. PROPOSITION. The unique n -form Λ of Proposition 18 is closed, $d\Lambda = 0$.

PROOF. Given $X_1, \dots, X_{n+1} \in M_p$, choose $u \in \varpi^{-1}(p)$ and $Y_1, \dots, Y_{n+1} \in \text{SO}(E)_u$ with $\varpi_* Y_i = X_i$. Let hY_i be the horizontal component of Y_i . Then

$$\begin{aligned} d\Lambda(X_1, \dots, X_{n+1}) &= d\Lambda(\varpi_* Y_1, \dots, \varpi_* Y_{n+1}) \\ &= d\Lambda(\varpi_* hY_1, \dots, \varpi_* hY_{n+1}) \\ &= (\varpi^* d\Lambda)(hY_1, \dots, hY_{n+1}) \\ &= d(\varpi^* \Lambda)(hY_1, \dots, hY_{n+1}) \\ &= 2^m m! d\{\text{Pf}(\Omega)\}(hY_1, \dots, hY_{n+1}) \\ &= 2^m m! D\{\text{Pf}(\Omega)\}(Y_1, \dots, Y_{n+1}). \end{aligned}$$

But $D\Omega = 0$ by Bianchi's identity (Theorem II.8-20), and this implies that $D\{\text{Pf}(\Omega)\} = 0$. \spadesuit

In view of Proposition 19, the n -form Λ determines a de Rham cohomology class $[\Lambda] \in H^n(M)$. The form Λ itself depends on the oriented n -dimensional bundle $\xi = \pi: E \rightarrow M$ over M , on the choice of a metric $\langle \cdot, \cdot \rangle$ for ξ , and on the connection ω on the corresponding bundle $\text{SO}(E)$.

20. PROPOSITION. The cohomology class $[\Lambda]$ is independent of the metric $\langle \cdot, \cdot \rangle$ and of the connection ω .

PROOF. Let $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle'$ be two metrics for ξ . By Corollary 5, the corresponding principle bundles $\text{SO}(E)$ and $\text{SO}'(E)$ are equivalent. If $\tilde{f}: \text{SO}'(E) \rightarrow \text{SO}(E)$ is a fibre preserving diffeomorphism which commutes with the action of $\text{SO}(n)$, and ω is a connection on $\text{SO}(E)$, then $\omega' = \tilde{f}^*\omega$ is a connection on $\text{SO}'(E)$. It is easy to see that the corresponding curvature forms satisfy $\Omega' = \tilde{f}^*\Omega$, so that $\text{Pf}(\Omega') = \tilde{f}^*\text{Pf}(\Omega)$. This implies that the corresponding forms Λ' and Λ are actually equal. To complete the proof it therefore suffices to show that any two connections ω_0, ω_1 on the same principal bundle $\text{SO}(E)$ give rise to forms Λ_0 and Λ_1 whose difference is exact.

Let $q: M \times [0, 1] \rightarrow M$ be the projection $q(p, t) = p$, and consider the bundle $q^*\text{SO}(\xi)$ over $M \times [0, 1]$. There are obvious induced connections $q^*\omega_0$ and $q^*\omega_1$ on $q^*\text{SO}(\xi)$. Let $\tau: M \times [0, 1] \rightarrow [0, 1]$ be the function $\tau(p, t) = t$, and form the connection

$$\omega = (1 - \tau)(q^*\omega_0) + \tau(q^*\omega_1)$$

on $q^*\text{SO}(\xi)$, with connection form Ω , say. If $i_t: M \rightarrow M \times [0, 1]$ is $i_t(p) = (p, t)$, then $i_0^*(\omega)$ can clearly be identified with ω_0 , and $i_1^*(\omega)$ can be identified with ω_1 . By Propositions 18 and 19 (which hold for manifolds-with-boundary as well as for manifolds), there is a closed n -form Λ on $M \times [0, 1]$ which pulls back to $2^m m! \text{Pf}(\Omega)$ on the total space of $q^*\text{SO}(\xi)$. Clearly we must have

$$i_0^*\Lambda = \Lambda_0 \quad \text{and} \quad i_1^*\Lambda = \Lambda_1.$$

Now Theorem I.7-17 (pg. I.224) shows that $\Lambda_1 - \Lambda_0$ is exact. ♦

We thus see that every oriented smooth bundle ξ over M of even fibre dimension n determines a de Rham cohomology class $C(\xi) = [\Lambda] \in H^n(N)$. Clearly $C(\xi) = C(\eta)$ if $\xi \simeq \eta$.

21. PROPOSITION. Let $\xi = \pi: E \rightarrow M$ be a smooth oriented bundle over M of even fibre dimension n , and let $f: M' \rightarrow M$ be a smooth map. Then

$$C(f^*\xi) = f^*(C(\xi)) \in H^n(M').$$

PROOF. Let E' be the total space of $f^*\xi$, and let $\tilde{f}: E' \rightarrow E$ be the bundle map covering f . If $\langle \cdot, \cdot \rangle$ is a metric on E , then $\tilde{f}^*\langle \cdot, \cdot \rangle$ is a metric on E' . Clearly there is an equivalence $\tilde{f}: \text{SO}(E') \rightarrow \text{SO}(E)$ covering f .

$$\begin{array}{ccc} \text{SO}(E') & \xrightarrow{\tilde{f}} & \text{SO}(E) \\ \varpi' \downarrow & & \downarrow \varpi \\ M' & \xrightarrow{f} & M \end{array}$$

If ω is a connection on $\text{SO}(E)$, then $\tilde{f}^*(\omega)$ will be a connection on $\text{SO}(E')$, and it is easy to see that the corresponding connection forms satisfy $\Omega' = \tilde{f}^*\Omega$. Consequently,

$$\text{Pf}(\Omega') = \text{Pf}(\tilde{f}^*\Omega) = \tilde{f}^*\text{Pf}(\Omega).$$

For the n -forms Λ on M given by Proposition 18 we then have

$$\begin{aligned} \varpi'^*(f^*\Lambda) &= \tilde{f}^*\varpi^*\Lambda \\ &= 2^m m! \tilde{f}^*\text{Pf}(\Omega) \\ &= 2^m m! \text{Pf}(\Omega'). \end{aligned}$$

So $f^*\Lambda$ must be the n -form Λ' on M' given by Proposition 18. ♦

We will extend the definition of C by setting $C(\xi) = 0$ when ξ is a smooth oriented bundle of odd fibre dimension. We would like to show that $C(\xi)$ is always some constant times $\chi(\xi)$. If this is the case, then we ought to have an analogue of Theorem 17 for C . And indeed we do.

22. THEOREM. Let $\xi_i = \pi_i: E_i \rightarrow M$ for $i = 1, 2$ be smooth oriented vector bundles over M , of fibre dimensions n_1 and n_2 . If $n_i = 2m_i$, then

$$C(\xi_1 \oplus \xi_2) = \frac{(m_1 + m_2)!}{m_1! m_2!} C(\xi_1) \cup C(\xi_2).$$

(For n_1 or n_2 odd, this just asserts that $C(\xi_1 \oplus \xi_2) = 0$.)

PROOF. Choose Riemannian metrics $\langle \cdot, \cdot \rangle_i$ on ξ_i , and let $\langle \cdot, \cdot \rangle$ be the obvious metric $\langle \cdot, \cdot \rangle_1 \oplus \langle \cdot, \cdot \rangle_2$ on $\xi_1 \oplus \xi_2 = \pi: E \rightarrow M$. Let $\varpi_i: \text{SO}(E_i) \rightarrow M$ and $\varpi: \text{SO}(E) \rightarrow M$ be the corresponding principal bundles. Over M we consider first the principal bundle $\text{SO}(E_1) * \text{SO}(E_2)$, with group $\text{SO}(n_1) \times \text{SO}(n_2) \subset \text{SO}(n_1 + n_2)$, whose fibre over $p \in M$ is just the direct product $\varpi_1^{-1}(p) \times \varpi_2^{-1}(p)$, so that we can regard

$$\text{SO}(E_1) * \text{SO}(E_2) \subset \text{SO}(E).$$

Let $\rho_i: \text{SO}(E_1) * \text{SO}(E_2) \rightarrow \text{SO}(E_i)$ be the obvious projection maps. If ω_i are connections on $\text{SO}(E_i)$, with curvature forms Ω_i , then

$$\rho_1^* \omega_1 \oplus \rho_2^* \omega_2 = \begin{pmatrix} \rho_1^* \omega_1 & 0 \\ 0 & \rho_2^* \omega_2 \end{pmatrix}$$

is a connection $\bar{\omega}$ on $\text{SO}(E_1) * \text{SO}(E_2)$, with curvature form

$$\bar{\Omega} = \rho_1^* \Omega_1 \oplus \rho_2^* \Omega_2 = \begin{pmatrix} \rho_1^* \Omega_1 & 0 \\ 0 & \rho_2^* \Omega_2 \end{pmatrix}.$$

The connection $\bar{\omega}$ can be extended uniquely to a connection $\tilde{\omega}$ on $\text{SO}(E)$ [the requirement $\tilde{\omega}(\sigma(M)) = M$ determines $\tilde{\omega}$ at the new vertical vectors, hence $\tilde{\omega}$ is determined at all points of $\text{SO}(E_1) * \text{SO}(E_2)$, and then at all points of $\text{SO}(E)$ by the requirement $\tilde{\omega}(R_A^* Y) = \text{Ad}(A^{-1})\tilde{\omega}(Y)$]. At any point $e \in \text{SO}(E_1) * \text{SO}(E_2)$ the horizontal vectors for $\tilde{\omega}$ are the same as for $\bar{\omega}$, so at e we have

$$\begin{aligned} \tilde{\Omega} &= \bar{\Omega} \quad [\text{on tangent vectors to } \text{SO}(E_1) * \text{SO}(E_2)] \\ \implies \text{Pf}(\tilde{\Omega}) &= \text{Pf}(\bar{\Omega}) = \text{Pf}(\rho_1^* \Omega_1) \wedge \text{Pf}(\rho_2^* \Omega_2) \quad \text{by Corollary 13} \\ &= \rho_1^* \text{Pf}(\Omega_1) \wedge \rho_2^* \text{Pf}(\Omega_2). \end{aligned}$$

So if Λ_i, Λ are the forms given by Proposition 18, then at e we must have

$$\begin{aligned} \varpi^* \Lambda &= 2^{m_1+m_2} (m_1 + m_2)! \text{Pf}(\tilde{\Omega}) \quad [\text{on tangent vectors to } \text{SO}(E_1) * \text{SO}(E_2)] \\ &= \frac{(m_1 + m_2)!}{m_1! m_2!} 2^{m_1} m_1! \rho_1^* \text{Pf}(\Omega_1) \wedge 2^{m_2} m_2! \rho_2^* \text{Pf}(\Omega_2) \\ &= \frac{(m_1 + m_2)!}{m_1! m_2!} \rho_1^* \varpi_1^* \Lambda_1 \wedge \rho_2^* \varpi_2^* \Lambda_2 \\ &= \frac{(m_1 + m_2)!}{m_1! m_2!} \varpi^* \Lambda_1 \wedge \varpi^* \Lambda_2. \end{aligned}$$

This implies that

$$\Lambda = \frac{(m_1 + m_2)!}{m_1! m_2!} \Lambda_1 \wedge \Lambda_2. \quad \spadesuit$$

Applying Theorem 22 when $n_1 = 1$, we immediately deduce that the class C has a property which we have already mentioned for χ (pg. I.445):

23. COROLLARY. If the oriented bundle $\xi = \pi: E \rightarrow M$ has a nowhere zero section s , then

$$C(\xi) = 0.$$

PROOF. Let $E_1 \subset E$ be

$$\bigcup_{p \in M} \mathbb{R} \cdot s(p),$$

and let $E_2 \subset E$ be the orthogonal complement

$$\bigcup_{p \in M} (\mathbb{R} \cdot s(p))^\perp$$

with respect to some Riemannian metric on E . Then $\xi_1 = \pi_1|E_1: E_1 \rightarrow M$ is an oriented 1-dimensional bundle, so $\xi_2 = \pi_2|E_2: E_2 \rightarrow M$ is also an oriented bundle (since ξ is oriented). Clearly $\xi \simeq \xi_1 \oplus \xi_2$. So Theorem 22 shows that $C(\xi) = 0$. ♦

But this fact practically characterizes χ :

24. COROLLARY. If $\xi = \pi: E \rightarrow M$ is a smooth oriented vector bundle of fibre dimension n over a compact oriented manifold M , then the class $C(\xi) \in H^n(M)$ is a multiple of the Euler class $\chi(\xi)$.

PROOF. Let S be the sphere bundle $S = \{e \in E : \langle e, e \rangle = 1\}$ formed with respect to some Riemannian metric on E , and let $\pi_0: S \rightarrow M$ be the restriction $\pi|S$. As we pointed out in section 2, the bundle $\pi_0^*\xi$ has a nowhere zero section. So Corollary 23 gives

$$\begin{aligned} 0 &= C(\pi_0^*\xi) \\ &= \pi_0^*C(\xi) \quad \text{by Proposition 21.} \end{aligned}$$

But Theorem I.11-31 says that a class $\alpha \in H^n(M)$ satisfies $\pi_0^*\alpha = 0$ if and only if α is a multiple of $\chi(\xi)$. ♦

If we apply this corollary to the tangent bundle of a compact oriented manifold M of even dimension n , we find that the class $C(TM) \in H^n(M)$ is some multiple of the Euler class $\chi(TM)$. This statement is not very interesting, since $H^n(M)$ is 1-dimensional (all it tells us is that $C(TM) = 0$ if $\chi(TM) = 0$). But we obtain a statement which is interesting when we apply the corollary to the universal bundle:

25. COROLLARY. For every even n , there is a “universal constant” A_n such that

$$C(\xi) = A_n \cdot \chi(\xi)$$

for *all* smooth oriented n -dimensional bundles ξ over compact oriented manifolds.

PROOF. Consider the bundles $\tilde{\gamma}^n(\mathbb{R}^N)$, for $N > n$. By Corollary 24, there are constants $A_{n,N}$ such that

$$(1) \quad C(\tilde{\gamma}^n(\mathbb{R}^N)) = A_{n,N} \cdot \chi(\tilde{\gamma}^n(\mathbb{R}^N)) \in H^n(\tilde{G}_n(\mathbb{R}^N)).$$

If $\alpha: \tilde{G}_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^M)$ is the natural inclusion, then

$$\alpha^*(\tilde{\gamma}^n(\mathbb{R}^M)) \simeq \tilde{\gamma}^n(\mathbb{R}^N),$$

so Propositions 15 and 21 give

$$(2) \quad C(\tilde{\gamma}^n(\mathbb{R}^N)) = \alpha^* C(\tilde{\gamma}^n(\mathbb{R}^M))$$

$$(3) \quad \chi(\tilde{\gamma}^n(\mathbb{R}^N)) = \alpha^* \chi(\tilde{\gamma}^n(\mathbb{R}^M)).$$

Equations (1)–(3) give

$$A_{n,N} \cdot \chi(\tilde{\gamma}^n(\mathbb{R}^N)) = A_{n,M} \cdot \chi(\tilde{\gamma}^n(\mathbb{R}^N)).$$

Since $\chi(\tilde{\gamma}^n(\mathbb{R}^N)) \neq 0$ by Corollary 16, this implies that $A_{n,N} = A_{n,M}$ for all $N, M > n$. Denoting this common number by A_n , we have

$$(*) \quad C(\tilde{\gamma}^n(\mathbb{R}^N)) = A_n \cdot \chi(\tilde{\gamma}^n(\mathbb{R}^N)).$$

Now by Theorem 8 any smooth oriented n -dimensional bundle ξ over a compact manifold M is equivalent to $f^*\tilde{\gamma}^n(\mathbb{R}^N)$ for some smooth map $f: M \rightarrow \tilde{G}_n(\mathbb{R}^N)$. Thus

$$\begin{aligned} C(\xi) &= C(f^*\tilde{\gamma}^n(\mathbb{R}^N)) \\ &= f^*C(\tilde{\gamma}^n(\mathbb{R}^N)) && \text{by Proposition 21} \\ &= A_n \cdot f^*\chi(\tilde{\gamma}^n(\mathbb{R}^N)) && \text{by } (*) \\ &= A_n \cdot \chi(\xi) && \text{by Proposition 15. } \spadesuit \end{aligned}$$

To see what this universal constant A_n is, we merely have to compute it in some convenient special case:

26. THEOREM (THE GAUSS-BONNET-CHERN THEOREM). For even $n = 2m$, the constant A_n is

$$\begin{aligned} A_n &= \frac{n!}{2} \cdot \text{volume of the unit } n\text{-sphere } S^n \\ &= \frac{n!}{2} \cdot \frac{\pi^m 2^{n+1} m!}{n!} = \pi^m 2^n m!. \end{aligned}$$

Consequently, if $(M, \langle \ , \ \rangle)$ is a compact oriented manifold of even dimension $n = 2m$, then

$$\begin{aligned} \int_M K_n dV &= \frac{1}{2} \text{volume of } S^n \cdot \chi(M) \\ &= \frac{\pi^m 2^n m!}{n!} \chi(M). \end{aligned}$$

PROOF. Let ξ be the tangent bundle TM of a compact oriented Riemannian n -manifold M . On page 284 we have a formula for $K_n dV$ (in this formula the Ω_j^i are the curvature forms for some positively oriented orthonormal moving frame) which clearly implies that the form Λ given by Proposition 18 for the bundle $\text{SO}(\xi) = \text{SO}(TM)$ is

$$\Lambda = n! K_n dV.$$

So if μ is the fundamental class of M , then

$$\begin{aligned} \left(\int_M K_n dV \right) \cdot \mu &= \frac{1}{n!} \left(\int_M \Lambda \right) \cdot \mu = \frac{1}{n!} C(\xi) \\ &= \frac{A_n}{n!} \chi(\xi) \\ &= \frac{A_n}{n!} \chi(M) \cdot \mu \quad \text{by Theorem I.11-30.} \end{aligned}$$

Hence

$$(1) \quad \int_M K_n dV = \frac{A_n}{n!} \chi(M).$$

Taking $M = S^n$ in (1), with $K_n = 1$, we have

$$\text{volume } S^n = \frac{A_n}{n!} \chi(S^n) = \frac{2A_n}{n!}$$

$$\implies A_n = \frac{n!}{2} \text{volume } S^n = \frac{n!}{2} \frac{\pi^m 2^{n+1} m!}{n!} = \pi^m 2^n m!, \quad \text{by Problem I.9-14.}$$

Substituting this value of A_n back into (1) we now have, for any M ,

$$\int_M K_n dV = \frac{\pi^m 2^n m!}{n!} \chi(M). \quad \spadesuit$$

5. THE CONCEPT OF CHARACTERISTIC CLASSES

Our proof of the generalized Gauss-Bonnet theorem made essential use of the fact that both the Euler class $\chi(\xi)$ and the class $C(\xi)$ are “natural”: for a bundle $\xi = \pi: E \rightarrow M$ and a map $f: M' \rightarrow M$ we have

$$\chi(f^*\xi) = f^*(\chi(\xi)) \quad \text{and} \quad C(f^*\xi) = f^*(C(\xi)).$$

This suggests that we might obtain greater insight into the theorem by trying to find out what *all* such natural classes are. To be precise, we define a **characteristic class of dimension k for smooth n -dimensional bundles** to be a function C which associates to each smooth n -dimensional bundle $\xi = \pi: E \rightarrow M$ an element

$$C(\xi) \in H^k(M),$$

with the following property: if $\xi' = \pi': E' \rightarrow M'$ is another smooth n -dimensional bundle, and (\tilde{f}, f) is a smooth bundle map from ξ' to ξ , then

$$C(\xi') = f^*(C(\xi)) \in H^n(M').$$

Here is an equivalent formulation: $C(\xi) = C(\eta)$ if $\xi \simeq \eta$, and for every smooth n -dimensional bundle $\xi = \pi: E \rightarrow M$ and smooth map $f: M' \rightarrow M$ we have

$$C(f^*\xi) = f^*(C(\xi)).$$

We can also define characteristic classes for oriented bundles; these are the characteristic classes that we will actually investigate. What we would like to do is to find out what *all* these characteristic classes are. This question might look hopeless, were it not for the universal bundles $\tilde{\gamma}^n(\mathbb{R}^N)$. Notice that a characteristic class C of dimension k for smooth n -dimensional bundles gives us, in particular, certain elements

$$c_N = C(\tilde{\gamma}^n(\mathbb{R}^N)) \in H^k(\tilde{G}_n(\mathbb{R}^N)).$$

If

$$\alpha_{N,N'}: \tilde{G}_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^{N'}) \quad N' > N$$

is the natural map, then

$$\alpha_{N,N'}^*(\tilde{\gamma}^n(\mathbb{R}^{N'})) \simeq \tilde{\gamma}^n(\mathbb{R}^N) \implies \alpha_{N,N'}^*c_{N'} = c_N.$$

Conversely, suppose we are given classes

$$c_N \in H^k(\tilde{G}_n(\mathbb{R}^N))$$

satisfying the compatibility condition

$$(C) \quad \alpha_{N,N'}^* c_{N'} = c_N \quad \text{for } N' > N.$$

Since, by Theorem 8, any oriented n -dimensional bundle ξ over a compact M is equivalent to $f^* \tilde{\gamma}^n(\mathbb{R}^N)$ for some $f: M \rightarrow \tilde{G}_n(\mathbb{R}^N)$, we can *define*

$$C(\xi) = f^* c_N.$$

Then $C(\xi)$ is well-defined, for if we have

$$f: M \rightarrow \tilde{G}_n(\mathbb{R}^N) \quad \text{and} \quad g: M \rightarrow \tilde{G}_n(\mathbb{R}^{N'})$$

with

$$f^* \tilde{\gamma}^n(\mathbb{R}^N) \simeq \xi \simeq g^* \tilde{\gamma}^n(\mathbb{R}^{N'}),$$

then the compositions

$$\alpha_{N,N''} \circ f, \quad \alpha_{N',N''} \circ g, \quad N'' \geq 2N, 2N'$$

are homotopic, so

$$\begin{aligned} f^* c_N &= f^* \alpha_{N,N''}^* c_{N''} && \text{by condition (C)} \\ &= (\alpha_{N,N''} \circ f)^* c_{N''} \\ &= (\alpha_{N',N''} \circ g)^* c_{N''} \\ &= g^* c_{N'}. \end{aligned}$$

Moreover, if $h: M' \rightarrow M$ is a smooth map, then

$$\begin{aligned} C(h^* \xi) &= C(h^* f^* \tilde{\gamma}^n(\mathbb{R}^N)) \\ &= h^* f^* c_N \\ &= h^* C(\xi). \end{aligned}$$

So we could just as well define a **characteristic class of dimension k for smooth oriented n -dimensional bundles** to be a collection of classes

$$c_N \in H^k(\tilde{G}_n(\mathbb{R}^N))$$

satisfying

$$(C) \quad \alpha_{N,N'}^* c_{N'} = c_N \quad \text{for } N' > N.$$

Now the problem doesn't seem quite so formidable; the main task seems to be the computation of $H^k(\tilde{G}_n(\mathbb{R}^N))$. As a matter of fact, it will turn out that the maps

$$\alpha_{N,N'}^*: H^k(\tilde{G}_n(\mathbb{R}^{N'})) \rightarrow H^k(\tilde{G}_n(\mathbb{R}^N))$$

are *isomorphisms* for $N, N' > n + k$. So a characteristic class of dimension k will be just the same as an element of $H^k(\tilde{G}_n(\mathbb{R}^N))$, for any $N > n + k$.

[If we were willing to work with singular cohomology, say, on spaces which are not manifolds, then we could define a characteristic class to be simply an element of the k -dimensional cohomology of the space $\tilde{G}_n(\mathbb{R}^\infty)$, the oriented version of the space $G_n(\mathbb{R}^\infty)$ defined on page 280.]

Now the calculation of cohomology groups is really the business of algebraic topologists, and all sorts of machinery has been used for computing characteristic classes [for all coefficient groups]. Rather than using any of the standard methods from this field, we will compute characteristic classes by purely differential geometric methods, making essential use of the fact that the Grassmannians are coset spaces of Lie groups. Although the procedure is quite involved, along the way we will get to look at several topics which are interesting in their own right. Moreover, the analysis will motivate the definition, in section 10, of one of the famous constructions in differential geometry. Finally, the Pfaffian will arise in a completely natural way.

6. THE COHOMOLOGY OF HOMOGENEOUS SPACES

By a **homogeneous space** we will mean a left coset space G/H , where G is a Lie group and H is a closed subgroup. We let $\pi: G \rightarrow G/H$ be the natural projection $\pi(a) = aH$, and we give G/H the quotient topology: a set $\mathcal{U} \subset G/H$ is open if and only if $\pi^{-1}(\mathcal{U}) \subset G$ is open. It is an easy exercise to show that G/H is Hausdorff. Notice also that if $V \subset G$ is any set, then

$$\begin{aligned} \pi(a) \in \pi(V) &\iff aH \in \{bH : b \in V\} \\ &\iff aH = bH && \text{for some } b \in V \\ &\iff a \in bH && \text{for some } b \in V. \end{aligned}$$

So

$$\pi^{-1}(\pi(V)) = V \cdot H = \bigcup_{h \in H} V \cdot h.$$

This shows that $\pi(V) \subset G/H$ is open if $V \subset G$ is open; thus π is an open map, as well as a continuous one.

For every $a \in G$ we have a map $\mathbf{L}_a: G/H \rightarrow G/H$ given by $\mathbf{L}_a(bH) = abH$ (the notation $L_a: G \rightarrow G$ will be reserved for the map given by $L_a(b) = ab$). Obviously the diagram

$$\begin{array}{ccc} G & \xrightarrow{\pi} & G/H \\ L_a \downarrow & & \downarrow \mathbf{L}_a \\ G & \xrightarrow{\pi} & G/H \end{array}$$

commutes, which implies that \mathbf{L}_a is continuous; more generally, it is easy to see that the map

$$G \times G/H \rightarrow G \quad \text{given by} \quad (a, bH) \mapsto abH$$

is continuous.

In this section we will show that G/H is a manifold, and we will find a method of computing the de Rham cohomology $H^*(G/H)$ when G is compact and connected. More precisely, we will reduce the determination of the de Rham cohomology to purely algebraic calculations involving the Lie algebras of G and H . In section 9 we will carry out a sufficient portion of the algebraic calculations for the Grassmannians $\tilde{G}_n(\mathbb{R}^N) = \text{SO}(N)/\text{SO}(n) \times \text{SO}(N-n)$ to determine all characteristic classes for oriented bundles.

We already know, from Theorem I.10-15, that the closed subgroup H of G is a Lie subgroup; in fact, there is a C^∞ structure on H , with the relative topology, that makes it a Lie subgroup of G .

27. PROPOSITION. Let G be a Lie group of dimension n , and H a closed subgroup of dimension d . Then G/H is a topological manifold of dimension $n-d$, and there is a unique C^∞ structure on G/H such that

- (i) $\pi: G \rightarrow G/H$ is C^∞
- (ii) For every point of G/H there is a neighborhood \mathcal{U} and a C^∞ section $s: \mathcal{U} \rightarrow G$ (a map $s: \mathcal{U} \rightarrow G$ satisfying $\pi \circ s = \text{identity}$).

PROOF. From the proof of Theorem I.10-15 we know that there is a coordinate system (x, U) around e with

$$\begin{aligned} x(e) &= 0 \\ x(U) &= (-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon), \end{aligned}$$

such that each slice

$$x^{d+1} = \text{constant}, \dots, x^n = \text{constant}$$

is an open subset of some left coset of H . In particular, the slice S_e through e is an open subset of H . Since H has the relative topology, this slice is of the form $V \cap H$ for some open set V . So by choosing ε smaller, if necessary, we can assume that

$$U \cap H = S_e.$$

We will now show that we can arrange for all slices to lie on different cosets of H . Choose $\varepsilon_1, \varepsilon_2 < \varepsilon$ so that the sets U_i with $x(U_i) = (-\varepsilon_i, \varepsilon_i) \times \cdots \times (-\varepsilon_i, \varepsilon_i)$ satisfy

$$U_1^{-1} \cdot U_1 \subset U_2, \quad U_2 \cdot U_2 \subset U.$$

If $a, b \in U_1$ satisfy $aH = bH$, then

$$b^{-1}a \in U_2 \cap H = U_2 \cap S_e \implies a \in b \cdot (U_2 \cap S_e).$$

Now $b \cdot (U_2 \cap S_e)$ is connected and lies in U , so it lies in a single slice. This shows that a and b lie in the same slice. Equivalently, different slices of U_1 lie in different cosets, as desired. For convenience, we assume that U_1 is our original U .

If we now consider the “cross-section” $C \subset U$ defined by

$$C = \{a \in U_1 : x^1(a) = \cdots = x^d(a) = 0\},$$

we see that

$$\pi|C : C \rightarrow G/H$$

is one-one, with image $\pi(C) = \pi(U)$, which is open in G/H . Since π is both continuous and open, the map $\pi|C$ is a homeomorphism. The inverse homeomorphism

$$\chi = (\pi|C)^{-1} : \pi(U) \rightarrow C$$

can be regarded as a map into \mathbb{R}^{n-d} ; we will use this map as a coordinate system around the coset H in G/H . For every $a \in G$ we let χ_a be the composition

$$\pi(a \cdot U) \xrightarrow{\mathbf{L}_{a^{-1}}} \pi(U) \xrightarrow{(\pi|C)^{-1}} C.$$

Then for $a, b \in G$ we have

$$\begin{aligned} \chi_a \circ \chi_b^{-1} &= (\pi|C)^{-1} \circ \mathbf{L}_{a^{-1}} \circ \mathbf{L}_b \circ \pi|C \\ &\text{on the set } W = \chi_b(\pi(a \cdot U) \cap \pi(b \cdot U)). \end{aligned}$$

For $c \in C$ we have

$$\mathbf{L}_{a^{-1}} \mathbf{L}_b(\pi|C)(c) = a^{-1}bcH,$$

and it is easy to check that if $c \in W$, then

$$(\pi|C)^{-1}L_{a^{-1}}L_b(\pi|C)(c) = a^{-1}bc.$$

This shows that χ_a and χ_b are C^∞ related, so the collection $\{\chi_a\}$ determines a C^∞ structure on G/H .

To show that $\pi: G \rightarrow G/H$ is C^∞ at a , we have to show that the map

$$(-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon) \xrightarrow{L_a \circ x^{-1}} a \cdot U \xrightarrow{\pi} \pi(a \cdot U) \xrightarrow{\chi_a} C$$

is C^∞ . This map is

$$(-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon) \xrightarrow{L_a \circ x^{-1}} a \cdot U \xrightarrow{\pi} \pi(a \cdot U) \xrightarrow{L_{a^{-1}}} \pi(U) \xrightarrow{(\pi|C)^{-1}} C,$$

which equals

$$(-\varepsilon, \varepsilon) \times \cdots \times (-\varepsilon, \varepsilon) \xrightarrow{x^{-1}} U \xrightarrow{\pi} \pi(U) \xrightarrow{(\pi|C)^{-1}} C;$$

the latter map is just projection on the last $n - d$ coordinates.

To prove (ii), we note that

$$s = L_a \circ \chi_a: \pi(a \cdot U) \rightarrow G$$

satisfies $\pi \circ s = \text{identity}$.

Uniqueness is left to the reader. ♦

The quotient topology on G/H has the property that $f: G/H \rightarrow X$ is continuous if and only if $f \circ \pi: G \rightarrow X$ is continuous. The C^∞ structure on G/H given by Proposition 27 now has the property that $f: G/H \rightarrow M$ is C^∞ if and only if $f \circ \pi: G \rightarrow M$ is C^∞ . In fact, if $f \circ \pi$ is C^∞ , and s is a C^∞ section on $\mathcal{U} \subset G/H$, then $f|_{\mathcal{U}} = f \circ \pi \circ s$ is C^∞ . It is also easy to see that the map

$$G \times G/H \rightarrow G/H, \quad (a, bH) \mapsto abH$$

is C^∞ : if $s: \mathcal{U} \rightarrow G$ is a section, then on $G \times \mathcal{U}$ this map equals

$$G \times \mathcal{U} \xrightarrow{\text{identity} \times s} G \times G \xrightarrow{\cdot} G \xrightarrow{\pi} G/H.$$

In particular, each $L_a: G/H \rightarrow G/H$ is C^∞ . Finally, we recall that in section 3 we defined a C^∞ structure on $\tilde{G}_n(\mathbb{R}^N)$ geometrically. It is easily checked that

this C^∞ structure satisfies (i) and (ii) when we consider $\tilde{G}_n(\mathbb{R}^N)$ as the quotient space $SO(N)/SO(n) \times SO(N-n)$.

For the remainder of this section we will assume that G is a *compact, connected* Lie group, with Lie algebra \mathfrak{g} , and that H is a closed subgroup with Lie algebra $\mathfrak{h} \subset \mathfrak{g}$. Before we consider the cohomology of G/H , a few preliminaries are needed. Recall (Proposition I.10-20) that any left invariant n -form σ^n on G is also right invariant. We will choose σ^n to be the unique bi-invariant n -form with $\int_G \sigma^n = 1$; as before, for a function $f: G \rightarrow \mathbb{R}$ we often write

$$\int_G f \sigma^n \quad \text{as} \quad \int_G f(a) da.$$

For every $a \in G$ we define the map $\text{Ad}(a): \mathfrak{g} \rightarrow \mathfrak{g}$ by

$$\text{Ad}(a) = (L_a \circ R_{a^{-1}})_*: \mathfrak{g} \rightarrow \mathfrak{g}.$$

When G is a subgroup of $GL(n, \mathbb{R})$, so that \mathfrak{g} is a subspace of $\mathfrak{gl}(n, \mathbb{R}) = n \times n$ matrices, we have the simple formula (Problem I.10-19 or pg. II.309)

$$\text{Ad}(A)M = AMA^{-1} \quad \text{for } A \in G, \quad M \in \mathfrak{g}.$$

Finally, recall that there is a bi-invariant Riemannian metric $\langle \cdot, \cdot \rangle$ on the compact Lie group G . When $G = O(n)$ we can describe such a metric explicitly as follows. For $M, P \in \mathfrak{o}(n) = \text{skew-symmetric } n \times n \text{ matrices}$, let

$$\langle M, P \rangle = \text{trace } MP^t = \sum_{i,j} M_{ij} P_{ij}.$$

For every $A \in O(n)$ we have

$$\begin{aligned} \langle \text{Ad}(A)M, \text{Ad}(A)P \rangle &= \text{trace } AMA^{-1} \cdot (APA^{-1})^t \\ &= \text{trace } A(MP^t)A^{-1} \\ &= \text{trace } MP^t = \langle M, P \rangle. \end{aligned}$$

So if we extend $\langle \cdot, \cdot \rangle$ to $O(n)$ by left invariance, then it will also be right invariant.

Now consider a k -form ω on G/H . We say that ω is **invariant** if $L_a^* \omega = \omega$ for all $a \in G$. For any k -form ω on G/H we can define a new k -form

$$\omega' = \int_G (a \mapsto L_a^* \omega) \sigma^n = \int_G L_a^* \omega da;$$

this equation really means that

$$\omega'(X_1, \dots, X_k) = \int_G [a \mapsto \mathbf{L}_a^* \omega(X_1, \dots, X_k)] \sigma^n = \int_G \mathbf{L}_a^* \omega(X_1, \dots, X_k) da.$$

It follows easily from left invariance of σ^n that ω' is invariant. More generally, given a smooth family $a \mapsto \eta_a$ of k -forms on G/H , where η_a is defined for all a in an open set $U \subset G$, we can form

$$\int_U (a \mapsto \eta_a) \sigma^n = \int_U \eta_a da,$$

which is a k -form on all of G/H . If X_1, \dots, X_k, Y are vector fields on G/H , then

$$Y \left(\int_U \eta_a(X_1, \dots, X_k) da \right) = \int_U Y(\eta_a(X_1, \dots, X_k)) da;$$

this follows from Proposition 9-10 when we choose an integral curve c for Y and let

$$\Gamma(u) = [a \mapsto \eta_a(X_1(c(u)), \dots, X_k(c(u)))] \cdot \sigma^n.$$

Using Theorem I.7-13, we then see that

$$d \left(\int_U (a \mapsto \eta_a) \sigma^n \right) = \int_U (a \mapsto d\eta_a) \sigma^n;$$

in simpler, but rather confusing, notation, we have

$$d \left(\int_U \eta_a da \right) = \int_U d\eta_a da.$$

28. PROPOSITION. If ω is a closed k -form on G/H , and

$$\omega' = \int_G \mathbf{L}_a^* \omega da,$$

then $\omega - \omega'$ is exact.

PROOF. For a fixed $a \in G$, the map $\mathbf{L}_a: G/H \rightarrow G/H$ is smoothly homotopic to the identity map $\mathbf{L}_e: G/H \rightarrow G/H$. In fact, we can write $a = \exp X$ for some $X \in \mathfrak{g}$ (Problem I.10-27), and consider the smooth family of maps $\mathbf{L}_{\exp tX}$ from \mathbf{L}_e to \mathbf{L}_a . It follows from Theorem I.8-13 that there is a $(k-1)$ -form η_a with

$$(*) \quad \omega - \mathbf{L}_a^* \omega = d\eta_a.$$

Moreover, the proof of this Theorem gives us an explicit formula for η_a .

Now let $E \subset \mathfrak{g}$ be an open set on which \exp is a diffeomorphism. From the explicit description of η_a we see that η_a varies smoothly with a for all $a \in \exp(E)$. So we can integrate equation (*) over $\exp(E)$, to obtain

$$[\text{volume } \exp(E)] \cdot \omega - \int_{\exp(E)} \mathbf{L}_a^* \omega \, da = d \left(\int_{\exp(E)} \eta_a \, da \right).$$

(Notice that the three forms in this equation are each defined on all of G/H .) But by Theorem 8-31, we can choose E so that $G - \exp(E)$ has measure 0. Then our equation becomes

$$\omega - \int_G \mathbf{L}_a^* \omega \, da = d \left(\int_{\exp(E)} \eta_a \, da \right). \spadesuit$$

On the other hand, it is even easier to show

29. PROPOSITION. If ω is a form on G/H which is invariant and exact, then ω is actually the exterior derivative of some invariant form.

PROOF. If $\omega = d\eta$, then

$$\begin{aligned} \omega &= \int_G \mathbf{L}_a^* \omega \, da = \int_G \mathbf{L}_a^* (d\eta) \, da \\ &= d \left(\int_G \mathbf{L}_a^* \eta \, da \right) = d\eta', \end{aligned}$$

where η' is invariant. \spadesuit

From Propositions 28 and 29 we immediately have

30. THEOREM. The k -dimensional de Rham cohomology $H^k(G/H)$ of G/H is naturally isomorphic to

$$H^k(G/H) \approx \frac{\text{closed invariant } k\text{-forms on } G/H}{\{d\eta : \eta \text{ an invariant } (k-1)\text{-form on } G/H\}}.$$

The cup product in $H^*(G/H)$ corresponds to \wedge under this isomorphism.

What really makes this result important is the fact that we can describe the invariant k -forms on G/H in terms of the left invariant forms on G . Note that the map $\text{Ad}(a): \mathfrak{g} \rightarrow \mathfrak{g}$ induces maps $\text{Ad}(a)^*: \Omega^k(\mathfrak{g}) \rightarrow \Omega^k(\mathfrak{g})$; the map $\text{Ad}(a)^*$ is just $(L_a \circ R_{a^{-1}})^*$ at e . A k -form η on G will be called **$\text{Ad}(H)$ -invariant** if

$$\text{Ad}(a)^*\eta(e) = \eta(e) \quad \text{for all } a \in H.$$

We say that η **annihilates** \mathfrak{h} if $\eta(e)(X_1, \dots, X_k) = 0$ whenever some $X_i \in \mathfrak{h}$.

31. LEMMA. If $\pi: G \rightarrow G/H$ is the natural projection, then the map $\omega \mapsto \pi^*\omega$ is a one-one correspondence between the invariant k -forms on G/H and the left invariant, $\text{Ad}(H)$ -invariant k -forms on G which annihilate \mathfrak{h} .

PROOF. If ω is invariant, then for all $a \in G$ we have

$$L_a^*\pi^*\omega = \pi^*L_a^*\omega = \pi^*\omega,$$

so $\pi^*\omega$ is left invariant. If $X_1, \dots, X_k \in \mathfrak{g}$ and some $X_i \in \mathfrak{h}$, then $\pi_*X_i = 0$, so

$$(\pi^*\omega)(X_1, \dots, X_k) = \omega(\pi_*X_1, \dots, \pi_*X_i, \dots, \pi_*X_k) = 0;$$

thus $\pi^*\omega$ annihilates \mathfrak{h} . Finally, if $a \in H$, then the map

$$\pi \circ L_a \circ R_{a^{-1}}: G \rightarrow G/H$$

is

$$b \mapsto aba^{-1}H = abH = L_a(bH) = L_a(\pi(b)),$$

so we have

$$(1) \quad \pi \circ L_a \circ R_{a^{-1}} = L_a \circ \pi;$$

consequently,

$$(L_a \circ R_{a^{-1}})^*\pi^*\omega = \pi^*L_a^*\omega = \pi^*\omega,$$

so $\pi^*\omega$ is $\text{Ad}(H)$ -invariant.

Conversely, suppose that the k -form η on G is left invariant, $\text{Ad}(H)$ -invariant, and annihilates \mathfrak{h} . The map $\pi_*: \mathfrak{g} \rightarrow (G/H)_H$ from \mathfrak{g} to the tangent space of G/H at the coset H has kernel precisely \mathfrak{h} , and therefore induces an isomorphism

$$\pi_*: \mathfrak{g}/\mathfrak{h} \rightarrow (G/H)_H.$$

We can consider $\eta(e) \in \Omega^k(\mathfrak{g}/\mathfrak{h})$, since η annihilates \mathfrak{h} , so there is a unique $\omega(H) \in \Omega^k((G/H)_H)$ with

$$(2) \quad \pi^*(\omega(H)) = \eta(e) \in \Omega^k(\mathfrak{g}/\mathfrak{h}).$$

Define ω on G/H by

$$\omega(aH) = L_{a^{-1}}^* \omega(H).$$

To show that ω is well-defined, consider $a, b \in G$ with $aH = bH$. Then $c = a^{-1}b \in H$. So by (1),

$$\begin{aligned} \pi^* L_c^* \omega(H) &= (L_c \circ R_{c^{-1}})^* \pi^* \omega(H) \\ &= (L_c \circ R_{c^{-1}})^* \eta(e) \\ &= \eta(e), \end{aligned}$$

since η is $\text{Ad}(H)$ -invariant. Since $\omega(H)$ is the unique element satisfying (2), we conclude that

$$\begin{aligned} \omega(H) &= L_c^* \omega(H) = (L_{a^{-1}} \circ L_b)^* \omega(H) \\ &= L_b^* (L_{a^{-1}}^* \omega(H)), \end{aligned}$$

and hence

$$L_{b^{-1}}^* \omega(H) = L_{a^{-1}}^* \omega(H),$$

as desired. Clearly ω is invariant. Moreover, for all $a \in G$ we have

$$\begin{aligned} \eta(a) &= L_{a^{-1}}^* \eta(e) \quad \text{since } \eta \text{ is left invariant} \\ &= L_{a^{-1}}^* \pi^* \omega(H) \\ &= \pi^* L_{a^{-1}}^* \omega(H) \\ &= \pi^*(\omega(aH)) \\ &= (\pi^* \omega)(a). \quad \blacklozenge \end{aligned}$$

Notice that in the first part of the proof we showed that for all $a \in H$, the given ω satisfies $(L_a \circ R_{a^{-1}})^* \pi^* \omega = \pi^* \omega$ at all points, not just at e . Since any left invariant form η on G which annihilates \mathfrak{h} and satisfies $(L_a \circ R_{a^{-1}})^* \eta(e) = \eta(e)$ is $\pi^* \omega$ for an invariant form ω on G/H , it follows that such an η satisfies $(L_a \circ R_{a^{-1}})^* \eta = \eta$ at all points, for all $a \in H$; this conclusion does not follow just from the fact that η is left invariant—we need to know that η annihilates \mathfrak{h} . Similarly, but more important, since $d\omega$ is clearly invariant for all invariant ω on G/H , it follows that if η is left invariant, $\text{Ad}(H)$ -invariant, and annihilates \mathfrak{h} ,

then $d\eta$ has these same properties [but the fact that η annihilates \mathfrak{h} , for example, does not by itself imply that $d\eta$ annihilates \mathfrak{h}].

We are now ready to give a completely algebraic description of $H^k(G/H)$. Notice first that if ω is a left invariant k -form, and $X_1, \dots, X_{k+1} \in \mathfrak{g}$, then $d\omega(X_1, \dots, X_{k+1})$ can be computed by applying Theorem I.7-13 to the left invariant vector fields \tilde{X}_i extending X_i . The terms

$$\omega(\tilde{X}_1, \dots, \widehat{\tilde{X}_i}, \dots, \tilde{X}_{k+1})$$

are all constant by left invariance, so our formula becomes simply

$$(*) \quad d\omega(X_1, \dots, X_{k+1}) = \sum_{i < j} (-1)^{i+j} \omega([X_i, X_j], X_1, \dots, \widehat{X_i}, \dots, \widehat{X_j}, \dots, X_{k+1}),$$

which involves only the bracket operation in \mathfrak{g} . Now let

$$\Omega^k(\mathfrak{g}/\mathfrak{h}) = \{\omega \in \Omega^k(\mathfrak{g}) : \omega(X_1, \dots, X_k) = 0 \text{ if some } X_i \in \mathfrak{h}\}.$$

The elements of $\Omega^k(\mathfrak{g}/\mathfrak{h})$ are clearly in one-one correspondence with the left invariant forms on G which annihilate \mathfrak{h} . If $a \in H$, and $X_i \in \mathfrak{h}$, then $\text{Ad}(a)X_i \in \mathfrak{h}$, since $L_a \circ R_{a^{-1}} : H \rightarrow H$; this shows that

$$\text{Ad}(a)^* : \Omega^k(\mathfrak{g}/\mathfrak{h}) \rightarrow \Omega^k(\mathfrak{g}/\mathfrak{h}) \quad \text{for all } a \in H.$$

Let

$$\Omega^k(\mathfrak{g}/\mathfrak{h})^H = \{\omega \in \Omega^k(\mathfrak{g}/\mathfrak{h}) : \text{Ad}(a)^*\omega = \omega \text{ for all } a \in H\}.$$

The remarks in the previous paragraph show that

$$d : \Omega^k(\mathfrak{g}/\mathfrak{h})^H \rightarrow \Omega^{k+1}(\mathfrak{g}/\mathfrak{h})^H,$$

where d is now defined by (*). Moreover, Theorem 30 shows that

$$H^k(G/H) \approx \frac{\ker d : \Omega^k(\mathfrak{g}/\mathfrak{h})^H \rightarrow \Omega^{k+1}(\mathfrak{g}/\mathfrak{h})^H}{d(\Omega^{k-1}(\mathfrak{g}/\mathfrak{h})^H)}.$$

Even this description of $H^k(G/H)$ can be simplified. Notice that if $\mathfrak{h}' \subset \mathfrak{g}$ is a subspace with $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{h}'$, then the elements of $\Omega^k(\mathfrak{h}')$ are in one-one

correspondence with the elements of $\Omega^k(\mathfrak{g}/\mathfrak{h})$: given $\omega \in \Omega^k(\mathfrak{h}')$, we define the corresponding $\bar{\omega} \in \Omega^k(\mathfrak{g}/\mathfrak{h})$ by

$$\bar{\omega}(X_1, \dots, X_k) = \omega(\mathfrak{h}' \text{ component of } X_1, \dots, \mathfrak{h}' \text{ component of } X_k).$$

In particular, consider the orthogonal complement $\mathfrak{h}^\perp \subset \mathfrak{g}$ of \mathfrak{h} with respect to $\langle \cdot, \cdot \rangle_e$, where $\langle \cdot, \cdot \rangle$ is a bi-invariant metric on G . Since each L_a and R_a is an isometry of $(G, \langle \cdot, \cdot \rangle)$, the map $\text{Ad}(a): \mathfrak{g} \rightarrow \mathfrak{g}$ is an isometry with respect to $\langle \cdot, \cdot \rangle_e$. Since $\text{Ad}(a): \mathfrak{h} \rightarrow \mathfrak{h}$ for $a \in H$, we also have $\text{Ad}(a): \mathfrak{h}^\perp \rightarrow \mathfrak{h}^\perp$ for $a \in H$, and hence

$$\text{Ad}(a)^*: \Omega^k(\mathfrak{h}^\perp) \rightarrow \Omega^k(\mathfrak{h}^\perp) \quad \text{for all } a \in H.$$

If we define

$$\Omega^k(\mathfrak{h}^\perp)^H = \{\omega \in \Omega^k(\mathfrak{h}^\perp) : \text{Ad}(a)^*\omega = \omega \text{ for all } a \in H\},$$

then the elements of $\Omega^k(\mathfrak{h}^\perp)^H$ are in one-one correspondence with the elements of $\Omega^k(\mathfrak{g}/\mathfrak{h})^H$. Putting all this information together, we have finally

32. THEOREM. Let G be a compact connected Lie group, and H a closed subgroup. Then the k -dimensional de Rham cohomology $H^k(G/H)$ of G/H is naturally isomorphic to

$$H^k(G/H) \approx \frac{\ker d: \Omega^k(\mathfrak{h}^\perp)^H \rightarrow \Omega^{k+1}(\mathfrak{h}^\perp)^H}{d(\Omega^{k-1}(\mathfrak{h}^\perp)^H)},$$

where

$\mathfrak{h}^\perp \subset \mathfrak{g}$ is the orthogonal complement of \mathfrak{h}
with respect to a bi-invariant metric,

and

$$\Omega^k(\mathfrak{h}^\perp)^H = \{\omega \in \Omega^k(\mathfrak{h}^\perp) : \text{Ad}(a)^*\omega = \omega \text{ for all } a \in H\},$$

and d is defined by

$$\begin{aligned} d\omega(X_1, \dots, X_{k+1}) \\ = \sum_{i < j} (-1)^{i+j} \omega(\mathfrak{h}^\perp \text{ component of } [X_i, X_j], X_1, \dots, \widehat{X_i}, \dots, \widehat{X_j}, \dots, X_{k+1}), \\ \text{for } X_1, \dots, X_{k+1} \in \mathfrak{h}^\perp. \end{aligned}$$

The cup product in $H^*(G/H)$ corresponds to \wedge under this isomorphism.

Naturally, the simplest applications of Theorem 32 will occur when H is a large subgroup, so that \mathfrak{h}^\perp is small. As an example, we consider $G = \mathrm{SO}(n+1)$, with $H = \mathrm{SO}(n) \subset \mathrm{SO}(n+1)$, so that $G/H = \tilde{G}_n(\mathbb{R}^{n+1}) \approx \tilde{G}_1(\mathbb{R}^{n+1})$ is S^n . In this case, where the geometry is so simple, it is easiest to use Theorem 30 directly. It tells us that in computing $H^k(S^n)$, it suffices to consider k -forms ω which are invariant under the action of $\mathrm{SO}(n+1)$. In particular, at any point $p \in S^n$, the function $\omega(p) \in \Omega^k(S^n_p)$ must be invariant under any linear transformation $A: S^n_p \rightarrow S^n_p$ which is special orthogonal with respect to the usual inner product on S^n_p . Now if $0 < k < n$ and $X_1, \dots, X_k \in S^n_p$ are orthonormal, then there is an A of this sort with

$$A(X_1) = X_2, \quad A(X_2) = X_1, \quad A(X_i) = X_i, \quad i = 3, \dots, k.$$

Consequently

$$\omega(p)(X_1, \dots, X_k) = \omega(p)(X_2, X_1, \dots, X_k),$$

so $\omega(p)(X_1, \dots, X_k) = 0$. This implies that $\omega(p) = 0$. Hence $H^k(S^n) = 0$ for $0 < k < n$. For $k = n$, we can choose $\omega(p)$ to be a multiple of the volume element $\sigma(p)$ of S^n at p . Since ω must also be invariant under special orthogonal maps taking p to any other point $q \in S^n$, we see that ω must be a constant multiple of the volume element σ . We have $d\sigma = 0$ automatically, and since there are no invariant $(k-1)$ -forms, we see that $H^n(S^n) \approx \mathbb{R}$.

It will also be instructive to see what happens when we do not rely on the geometry, and use Theorem 32. The Lie algebra $\mathfrak{g} = \mathfrak{o}(n+1)$ is the set of all skew-symmetric $(n+1) \times (n+1)$ matrices, while \mathfrak{h} consists of those of the form

$$\begin{pmatrix} \boxed{M} & 0 \\ & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \quad M \in \mathfrak{o}(n).$$

For the bi-invariant metric $\langle M, N \rangle = \mathrm{trace} \, MN^t = \sum_{i,j} M_{ij} N_{ij}$, the orthogonal complement \mathfrak{h}^\perp is spanned by the n matrices

$$Y_1 = \begin{pmatrix} \boxed{0} & 1 \\ & 0 \\ & \vdots \\ -1 & 0 & \dots & 0 \end{pmatrix}, \quad \dots, \quad Y_n = \begin{pmatrix} \boxed{0} & 0 \\ & \vdots \\ & 1 \\ 0 & \dots & -1 & 0 \end{pmatrix}$$

For any matrix $\tilde{A} \in H$ of the form

$$\tilde{A} = \begin{pmatrix} \boxed{A} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 & \dots & 0 & 1 \end{pmatrix}, \quad A \in \mathrm{SO}(n),$$

we compute that

$$\begin{aligned} \tilde{A}Y_i\tilde{A}^{-1} &= \begin{pmatrix} \boxed{A} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} \boxed{0} & \begin{matrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{matrix} \\ 0 & \dots & -1 & \dots & 0 \end{pmatrix} \begin{pmatrix} \boxed{A^t} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 & \dots & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \boxed{A} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ 0 & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} \boxed{0} & \begin{matrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{matrix} \\ -a_{1i} & \dots & -a_{ni} & 0 \end{pmatrix} \\ &= \begin{pmatrix} \boxed{0} & \begin{matrix} a_{1i} \\ \vdots \\ a_{ni} \end{matrix} \\ -a_{1i} & \dots & -a_{ni} & 0 \end{pmatrix} \\ &= \sum_{j=1}^n a_{ji} Y_j. \end{aligned}$$

So if we regard the Y_i simply as vectors in \mathbb{R}^n , then the adjoint action $\mathrm{Ad}(\tilde{A})$ on the Y_i is just the usual action of the orthogonal matrix A on the vectors Y_i . As we saw in the previous paragraph, this means that

$$\Omega^k(\mathfrak{h}^\perp)^H = \begin{cases} 0 & 0 < k < n \\ \mathbb{R} & k = n; \end{cases}$$

hence $H^k(G/H) = 0$ for $0 < k < n$, and $H^n(G/H) \approx \mathbb{R}$.

7. A SMATTERING OF CLASSICAL INVARIANT THEORY

The simple algebraic considerations used at the conclusion of the previous section won't get us very far when we replace the subgroup $H = \text{SO}(n) \subset \text{SO}(n+1)$ by a smaller subgroup. In order to analyze the more general situation in an effective way, we need to delve briefly into classical invariant theory, which was once considered the cornerstone of all mathematics, and then rapidly dwindled to a state of near extinction, although recently it has excited new interest.

As an example of the sort of question that arises in invariant theory, we consider a standard fact from algebra, to which we have already alluded on occasion. A function

$$f: \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_m \rightarrow \mathbb{R}$$

is **symmetric** if

$$f(x_1, \dots, x_m) = f(x_{\pi(1)}, \dots, x_{\pi(m)})$$

for all permutations $\pi \in S_m$. Alternatively, if we define an operation of S_m on $\mathbb{R} \times \cdots \times \mathbb{R}$ by

$$\pi \cdot (x_1, \dots, x_m) = (x_{\pi(1)}, \dots, x_{\pi(m)}),$$

then f is symmetric if and only if

$$f(\pi \cdot x) = f(x) \quad \text{for all } x \in \mathbb{R} \times \cdots \times \mathbb{R} \quad \text{and all } \pi \in S_m.$$

As examples of symmetric functions we have the “elementary symmetric functions”

$$\begin{aligned} \sigma_1(x_1, \dots, x_m) &= \sum_{i=1}^m x_i, & \sigma_2(x_1, \dots, x_m) &= \sum_{i < j} x_i x_j \\ &\dots & & \\ \sigma_m(x_1, \dots, x_m) &= x_1 \cdots x_m; \end{aligned}$$

and for all $x, y \in \mathbb{R} \times \cdots \times \mathbb{R}$ we have $y = \pi \cdot x$ for some $\pi \in S_m$ if and only if $\sigma_i(x) = \sigma_i(y)$ for all i (compare pg. IV.65). From this we see immediately that any symmetric f can be written

$$f(x_1, \dots, x_m) = F(\sigma_1(x_1, \dots, x_m), \dots, \sigma_m(x_1, \dots, x_m))$$

for some function F . Indeed, we can define

$$F(s_1, \dots, s_m) = f(x_1, \dots, x_m) \quad \text{for any } x_1, \dots, x_m \text{ with } \sigma_i(x_1, \dots, x_m) = s_i$$

—such x_1, \dots, x_m certainly exist: we can take x_1, \dots, x_m to be the roots of the polynomial

$$x^n - s_1 x^{n-1} + \dots + (-1)^n s_n = 0.$$

On the other hand, if f is a *polynomial*, then it is by no means so evident that we can choose F to be a polynomial; the argument which establishes this fact involves a slightly delicate induction, and can be found in any standard algebra course.

Note, by the way, that the polynomials $\sigma_1, \dots, \sigma_m$ are algebraically independent, i.e., if p is any polynomial with

$$p(\sigma_1(x_1, \dots, x_m), \dots, \sigma_m(x_1, \dots, x_m)) = 0$$

for all x_1, \dots, x_m , then $p = 0$. In fact, this equation implies that

$$p(s_1, \dots, s_m) = 0$$

for all s_1, \dots, s_m , and hence that $p = 0$.

Now consider a function

$$f: \underbrace{\mathbb{R}^n \times \dots \times \mathbb{R}^n}_{m \text{ times}} \rightarrow \mathbb{R},$$

which we will often describe as a “function of m vectors in \mathbb{R}^n ”. A typical element of $\mathbb{R}^n \times \dots \times \mathbb{R}^n$ will be an m -tuple of vectors (v_1, \dots, v_m) , and each v_r is an n -tuple v_{r1}, \dots, v_{rn} . We say that a function f of m vectors in \mathbb{R}^n is **invariant under $O(n)$** if

$$f(v_1, \dots, v_m) = f(A(v_1), \dots, A(v_m))$$

for all $v_1, \dots, v_m \in \mathbb{R}^n$ and all $A \in O(n)$. Alternatively, if we define an action of $O(n)$ on $\mathbb{R}^n \times \dots \times \mathbb{R}^n$ by

$$A \cdot (v_1, \dots, v_m) = (A(v_1), \dots, A(v_m)),$$

then f is invariant under $O(n)$ if and only if

$$f(A \cdot v) = f(v) \quad \text{for all } v \in \mathbb{R}^n \times \dots \times \mathbb{R}^n \text{ and all } A \in O(n).$$

Similarly, we can consider functions invariant under any subgroup of $GL(n, \mathbb{R})$. In writing $A(v_r)$, we are considering an $n \times n$ matrix $A = (a_{ij})$ as a linear transformation $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$, by the rule

$$A(e_i) = \sum_{j=1}^n a_{ji} e_j.$$

Since we are regarding $v_r = (v_{r1}, \dots, v_{rn}) = \sum_h v_{rh} e_h$ as a row vector, this means that we have

$$\begin{aligned} A(v_r) &= \sum_h v_{rh} A(e_h) = \sum_{h,j} v_{rh} a_{jh} e_j \\ &= \left(\sum_h v_{rh} a_{1h}, \dots, \sum_h v_{rh} a_{nh} \right), \end{aligned}$$

and consequently

$$A(v_r) = v_r \cdot A^t \quad \begin{array}{l} \text{[the product of the } 1 \times n \\ \text{matrix } v_r \text{ with the } n \times n \\ \text{matrix } A^t], \end{array}$$

which is slightly unpleasant, but something we can live with.

If $v, w \in \mathbb{R}^n \times \dots \times \mathbb{R}^n$ are m -tuples of vectors with $\langle v_r, v_s \rangle = \langle w_r, w_s \rangle$ for all r, s , then there is $A \in O(n)$ with $w = A \cdot v$. It follows immediately that every function f of m vectors in \mathbb{R}^n which is invariant under $O(n)$ can be written as

$$f(v_1, \dots, v_m) = F(\langle v_1, v_1 \rangle, \dots, \langle v_m, v_m \rangle)$$

for some function F . For brevity, we will also write

$$f(v_1, \dots, v_m) = F(\{\langle v_r, v_s \rangle\}),$$

and if we introduce the inner product functions

$$\iota_{rs}(v_1, \dots, v_m) = \langle v_r, v_s \rangle,$$

then we can write

$$f = F \circ (\{\iota_{rs}\}).$$

From this general, and trivial result, however, it does not follow that every *polynomial* function of m vectors in \mathbb{R}^n can be written as a *polynomial* in the ι_{rs} (a function $f: \mathbb{R}^n \times \dots \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a **polynomial** function if $f(\sum a_1^i e_i, \dots, \sum a_m^i e_i)$ is a polynomial in the a_j^i). This deeper algebraic result is the content of the “first main theorem of invariant theory for $O(n)$ ”. In order to prove this result, as well as the corresponding result for $SO(n)$, we will follow the classical route, which will get us to our destination in the shortest time, although it involves some unpleasant calculations, and uses some mysterious identities.

First, some preliminaries about polynomial functions f of m vectors in \mathbb{R}^n . We say that f is **homogeneous of degree** $(\alpha_1, \dots, \alpha_m)$ if

$$f(\lambda_1 v_1, \dots, \lambda_m v_m) = \lambda_1^{\alpha_1} \dots \lambda_m^{\alpha_m} f(v_1, \dots, v_m).$$

Every polynomial function f can be written

$$(*) \quad f = \sum_{(\alpha_1, \dots, \alpha_m)} f_{\alpha_1, \dots, \alpha_m},$$

where $f_{\alpha_1, \dots, \alpha_m}$ is homogeneous of degree $(\alpha_1, \dots, \alpha_m)$. For example, we might have $f: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ with

$$f(x_1, x_2, y_1, y_2) = \underbrace{3x_1^2 x_2 y_1^2 + 7x_1 x_2^2 y_1 y_2}_{\text{degree } (3,2)} + \underbrace{8x_1^3}_{\text{degree } (3,0)} + \underbrace{7x_2^4}_{\text{degree } (4,0)}.$$

It is easy to see that the expression $(*)$ is unique. Moreover, if f is homogeneous of degree $(\alpha_1, \dots, \alpha_m)$, then so is

$$(v_1, \dots, v_m) \mapsto f(A \cdot v_1, \dots, A \cdot v_m)$$

for any linear transformation A . From this we easily see that if f is invariant under any group of linear transformations, then so is $f_{\alpha_1, \dots, \alpha_m}$. So we will henceforth consider only homogeneous polynomial functions.

Notice that if $f: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies $f(\lambda v) = \lambda^k f(v)$, then

$$k\lambda^{k-1} f(v) = \frac{d}{d\lambda} \lambda^k f(v) = \frac{d}{d\lambda} f(\lambda v) = \sum_{i=1}^n v_i \frac{\partial f}{\partial x_i}(\lambda v).$$

In particular, for $\lambda = 1$ we obtain *Euler's Theorem*

$$kf = \sum_{i=1}^n v_i \frac{\partial f}{\partial x_i}.$$

(In the case of a polynomial function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, this result can be verified directly.) Naturally, there is an analogous result for homogeneous functions of several vectors in \mathbb{R}^n .

Now let

$$\begin{array}{c} e_{ri} = (0, \dots, e_i, \dots, 0) \in \mathbb{R}^n \times \dots \times \mathbb{R}^n, \\ \uparrow \\ r^{\text{th}} \text{ place} \end{array}$$

where 0 denotes the zero vector of \mathbb{R}^n , and e_i is the i^{th} standard basis vector of \mathbb{R}^n . The e_{ri} form the standard basis for $\mathbb{R}^n \times \dots \times \mathbb{R}^n$ when we identify it with \mathbb{R}^{nm} , so for a function

$$f: \underbrace{\mathbb{R}^n \times \dots \times \mathbb{R}^n}_{m \text{ times}} \rightarrow \mathbb{R}$$

we can consider the partial derivatives

$$\frac{\partial f}{\partial e_{ri}};$$

these partial derivatives certainly exist if f is a polynomial function. Now for $1 \leq r, s \leq m$ we can consider the function

$$(D_{sr}f)(v_1, \dots, v_m) = \sum_{i=1}^n v_{si} \frac{\partial f}{\partial e_{ri}}(v_1, \dots, v_m);$$

in terms of the dual basis $\{\phi^{ri}\}$ to the $\{e_{ri}\}$, we can write

$$D_{sr} = \sum_{i=1}^n \phi^{si} \cdot \frac{\partial}{\partial e_{ri}}.$$

For example, if $(x_1, \dots, x_n, y_1, \dots, y_n)$ denotes a typical element of $\mathbb{R}^n \times \mathbb{R}^n$, then for $f: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ we have

$$D_{21}f(x_1, \dots, x_n, y_1, \dots, y_n) = \sum_{i=1}^n y_i \frac{\partial f}{\partial x_i}(x_1, \dots, x_n, y_1, \dots, y_n).$$

The operator D_{sr} is called a **polarization**. It is important for the following reason.

33. LEMMA. Suppose that

$$f(v_1, \dots, v_m) = f(A(v_1), \dots, A(v_m))$$

for all $v_1, \dots, v_m \in \mathbb{R}^n$ and some linear transformation A . Then also

$$D_{sr}f(v_1, \dots, v_m) = D_{sr}f(A(v_1), \dots, A(v_m))$$

for all $v_1, \dots, v_m \in \mathbb{R}^n$.

PROOF. If (a_{ij}) is the matrix of A , then by hypothesis we have

$$f(\dots, (v_{r1}, \dots, v_{rn}), \dots) = f(\dots, (\sum_{j=1}^n a_{1j} v_{rj}, \dots, \sum_{j=1}^n a_{nj} v_{rj}), \dots),$$

which implies that

$$\frac{\partial f}{\partial e_{ri}}(v_1, \dots, v_m) = \sum_{k=1}^n \frac{\partial f}{\partial e_{rk}}(\dots) a_{ki}.$$

Therefore

$$\begin{aligned} \sum_{i=1}^n v_{si} \frac{\partial f}{\partial e_{ri}}(v_1, \dots, v_m) &= \sum_{k=1}^n \sum_{i=1}^n a_{ki} v_{si} \frac{\partial f}{\partial e_{rk}}(\dots) \\ &= \sum_{k=1}^n A(v_s)_k \frac{\partial f}{\partial e_{rk}}(A(v_1), \dots, A(v_m)). \quad \spadesuit \end{aligned}$$

In particular, we see that polarization takes polynomial functions invariant under a group of matrices into polynomial functions with the same property. Note that if f is homogeneous of degree $(\alpha_1, \dots, \alpha_m)$, then Euler's theorem implies that $D_{rr} f = \alpha_r \cdot f$. Note also that polarizations take the inner product functions into sums of such:

$$D_{sr} l_{qp} = \delta_{rq} l_{sq} + \delta_{rp} l_{sp}.$$

Finally, consider a determinant function

$$(v_1, \dots, v_m) \mapsto \det \begin{pmatrix} v_{r_1} \\ \vdots \\ v_{r_n} \end{pmatrix},$$

which we will denote by $\det_{r_1 \dots r_n}$. We clearly have

$$\begin{aligned} D_{sr_i} \det_{r_1 \dots r_n} &= \det_{r_1 \dots r_{i-1} s r_{i+1} \dots r_n} \\ D_{sr} \det_{r_1 \dots r_n} &= 0 \quad \text{if } r \neq r_1, \dots, r_n. \end{aligned}$$

Now we want to look at the result of composing two or more polarizations. It would be nice if

$$D_{s_2 r_2} D_{s_1 r_1} f \stackrel{(?)}{=} \sum_{i_1, i_2} \phi^{s_1 i_1} \phi^{s_2 i_2} \frac{\partial^2 f}{\partial e_{r_1 i_1} \partial e_{r_2 i_2}}.$$

But this holds only when $r_2, s_2 \neq s_1$. It will be convenient, however, to denote the right side of the above equation by a symbol that looks like a composition, even in the case where r_2 or s_2 equals s_1 . So we will use the symbol Δ_{sr} for the same operator as D_{sr} , but we will define the operator $\Delta_{s_2 r_2} \Delta_{s_1 r_1}$ not as a composition, but formally by

$$\Delta_{s_2 r_2} \Delta_{s_1 r_1} = \sum_{i_1, i_2} \phi^{s_1 i_1} \phi^{s_2 i_2} \frac{\partial^2}{\partial e_{r_1 i_1} \partial e_{r_2 i_2}};$$

the operators $\Delta_{s_3 r_3} \Delta_{s_2 r_2} \Delta_{s_1 r_1}$, etc., are defined similarly. As for the actual composition $D_{s_2 r_2} D_{s_1 r_1}$ we have

$$\begin{aligned} D_{s_2 r_2} D_{s_1 r_1} f &= \sum_{i_2} \phi^{s_2 i_2} \frac{\partial}{\partial e_{r_2 i_2}} \left(\sum_{i_1} \phi^{s_1 i_1} \frac{\partial f}{\partial e_{r_1 i_1}} \right) \\ &= \sum_{i_1, i_2} \phi^{s_2 i_2} \phi^{s_1 i_1} \frac{\partial^2 f}{\partial e_{r_2 i_2} \partial e_{r_1 i_1}} + \delta_{s_1}^{r_2} \sum_{i_2} \phi^{s_2 i_2} \frac{\partial f}{\partial e_{r_1 i_2}}, \end{aligned}$$

which shows that

$$(1) \quad D_{s_2 r_2} D_{s_1 r_1} = \Delta_{s_2 r_2} \Delta_{s_1 r_1} + \delta_{s_1}^{r_2} \Delta_{s_2 r_1}.$$

Now consider the operator

$$\det \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \stackrel{\text{def}}{=} D_{11} D_{22} - D_{21} D_{12}.$$

From (1) we have

$$D_{11} D_{22} = \Delta_{11} \Delta_{22}, \quad D_{21} D_{12} = \Delta_{21} \Delta_{12} + \Delta_{22},$$

so

$$D_{11} D_{22} - D_{21} D_{12} + D_{22} = \Delta_{11} \Delta_{22} - \Delta_{21} \Delta_{12}.$$

We used the indices 1 and 2 for convenience, but we clearly have the same result for any distinct indices α_1, α_2 . We can write our equation as

$$(2) \quad \det \begin{pmatrix} D_{\alpha_1 \alpha_1} + 1 & D_{\alpha_1 \alpha_2} \\ D_{\alpha_2 \alpha_1} & D_{\alpha_2 \alpha_2} \end{pmatrix} = \det \begin{pmatrix} \Delta_{\alpha_1 \alpha_1} & \Delta_{\alpha_1 \alpha_2} \\ \Delta_{\alpha_2 \alpha_1} & \Delta_{\alpha_2 \alpha_2} \end{pmatrix}.$$

Remarkably enough, this equation can be generalized. First we compute that

$$\begin{aligned} D_{s_3 r_3} (\Delta_{s_2 r_2} \Delta_{s_1 r_1}) f &= \sum_{i_3} \phi^{s_3 i_3} \frac{\partial}{\partial e_{r_3 i_3}} \left(\sum_{i_1, i_2} \phi^{s_2 i_2} \phi^{s_1 i_1} \frac{\partial^2 f}{\partial e_{r_2 i_2} \partial e_{r_1 i_1}} \right) \\ &= \sum_{i_1, i_2, i_3} \phi^{s_3 i_3} \phi^{s_2 i_2} \phi^{s_1 i_1} \frac{\partial^3 f}{\partial e_{r_3 i_3} \partial e_{r_2 i_2} \partial e_{r_1 i_1}} \\ &\quad + \delta_{r_3}^{s_1} \sum_{i_2, i_3} \phi^{s_3 i_3} \phi^{s_2 i_2} \frac{\partial^2 f}{\partial e_{r_2 i_2} \partial e_{r_1 i_3}} \\ &\quad + \delta_{r_3}^{s_2} \sum_{i_1, i_3} \phi^{s_3 i_3} \phi^{s_1 i_1} \frac{\partial^2 f}{\partial e_{r_2 i_3} \partial e_{r_1 i_1}} \end{aligned}$$

(this formula works even if $s_1 = s_2$). Consequently,

$$(3) \quad D_{s_3 r_3} \circ (\Delta_{s_2 r_2} \Delta_{s_1 r_1}) = \Delta_{s_3 r_3} \Delta_{s_2 r_2} \Delta_{s_1 r_1} + \delta_{r_3}^{s_1} \Delta_{s_3 r_1} \Delta_{s_2 r_2} + \delta_{r_3}^{s_2} \Delta_{s_3 r_2} \Delta_{s_1 r_1}.$$

Now we consider the operator

$$\det \begin{pmatrix} D_{11} & \Delta_{12} & \Delta_{13} \\ D_{21} & \Delta_{22} & \Delta_{23} \\ D_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix} \stackrel{\text{def}}{=} \sum_{\pi \in \mathcal{S}_3} (\text{sgn } \pi) \cdot D_{\pi(1),1} \circ (\Delta_{\pi(2),2} \Delta_{\pi(3),3}).$$

[For each term of this sum, operators in the first column of the matrix appear on the left, followed by operators from the second column, etc.] Using (3) we have

$$\begin{aligned} & \det \begin{pmatrix} D_{11} & \Delta_{12} & \Delta_{13} \\ D_{21} & \Delta_{22} & \Delta_{23} \\ D_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix} \\ &= \sum_{\pi \in \mathcal{S}_3} (\text{sgn } \pi) \{ \Delta_{\pi(1),1} \Delta_{\pi(2),2} \Delta_{\pi(3),3} \\ & \quad + \delta_{\pi(3)}^1 \Delta_{\pi(1),3} \Delta_{\pi(2),2} \\ & \quad + \delta_{\pi(2)}^1 \Delta_{\pi(1),2} \Delta_{\pi(3),3} \} \\ &= \sum_{\pi \in \mathcal{S}_3} (\text{sgn } \pi) \{ \Delta_{\pi(1),1} \Delta_{\pi(2),2} \Delta_{\pi(3),3} \\ & \quad - \delta_{\pi(1)}^1 \Delta_{\pi(3),3} \Delta_{\pi(2),2} \quad \begin{array}{l} \text{[compose } \pi \text{ with the trans-} \\ \text{position interchanging} \\ \pi(1) \text{ and } \pi(3)] \end{array} \\ & \quad - \delta_{\pi(1)}^1 \Delta_{\pi(2),2} \Delta_{\pi(3),3} \} \quad \begin{array}{l} \text{[compose } \pi \text{ with the trans-} \\ \text{position interchanging} \\ \pi(1) \text{ and } \pi(2)] \end{array} \\ &= \sum_{\pi \in \mathcal{S}_3} (\text{sgn } \pi) \Delta_{\pi(1),1} \Delta_{\pi(2),2} \Delta_{\pi(3),3} \\ & \quad - 2 \sum_{\substack{\pi \in \mathcal{S}_3 \\ \text{with} \\ \pi(1)=1}} (\text{sgn } \pi) \Delta_{\pi(2),2} \Delta_{\pi(3),3} \\ &= \det \begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix} - 2 \det \begin{pmatrix} \Delta_{22} & \Delta_{23} \\ \Delta_{32} & \Delta_{33} \end{pmatrix}. \end{aligned}$$

We can also write this equation as

$$\det \begin{pmatrix} D_{11} + 2 & \Delta_{12} & \Delta_{13} \\ D_{21} & \Delta_{22} & \Delta_{23} \\ D_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix} = \det \begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix}.$$

Using (2), we find that

$$\det \begin{pmatrix} D_{11} + 2 & D_{12} & D_{13} \\ D_{21} & D_{22} + 1 & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix} = \det \begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} \end{pmatrix}.$$

Of course, the numbers 1, 2, 3 could be replaced by any three distinct integers $\alpha_1, \alpha_2, \alpha_3$ from 1 to m . The same general procedure yields, by induction, the result

$$(4) \quad \det \begin{pmatrix} D_{11} + (m-1) & D_{12} & \dots & D_{1m} \\ \vdots & D_{22} + (m-2) & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ D_{m1} & D_{m2} & \dots & D_{mm} \end{pmatrix} = \det \begin{pmatrix} \Delta_{11} & \dots & \Delta_{1m} \\ \vdots & & \vdots \\ \Delta_{m1} & \dots & \Delta_{mm} \end{pmatrix}.$$

We introduce the **Cayley Ω -process** which takes a function f of n vectors in \mathbb{R}^n to the function Ωf of n vectors defined by

$$\Omega f = \det \begin{pmatrix} \frac{\partial}{\partial e_{11}} & \dots & \frac{\partial}{\partial e_{1n}} \\ \vdots & & \vdots \\ \frac{\partial}{\partial e_{n1}} & \dots & \frac{\partial}{\partial e_{nn}} \end{pmatrix} f.$$

Notice that we could just as well write the transpose matrix here, since all partials commute. It is easily seen that

$$\Omega f(A \cdot v_1, \dots, A \cdot v_n) = (\det A) \cdot \Omega f(v_1, \dots, v_n).$$

So if f is invariant under $O(n)$ or $SO(n)$, then Ωf is invariant under $SO(n)$. Using \det for the function

$$(v_1, \dots, v_n) \mapsto \det \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

we now have

34. THEOREM (THE CAPELLI IDENTITIES). Let f be a polynomial function of m vectors in \mathbb{R}^n . Then

$$\det \begin{pmatrix} D_{11} + (m-1) & \dots & D_{1m} \\ \vdots & \ddots & \vdots \\ D_{m1} & \dots & D_{mm} \end{pmatrix} f = \begin{cases} 0 & m > n \\ \det \Omega f & m = n. \end{cases}$$

PROOF. Equation (4) shows that at (v_1, \dots, v_m) the left side of our equation has the value

$$\begin{aligned} & \sum_{\pi \in S_m} (\operatorname{sgn} \pi) \Delta_{\pi(1),1} \cdots \Delta_{\pi(m),m} f(v_1, \dots, v_m) \\ &= \sum_{\pi \in S_m} (\operatorname{sgn} \pi) \sum_{i_1, \dots, i_m=1}^n v_{\pi(1),i_1} \cdots v_{\pi(m),i_m} \frac{\partial^m f}{\partial e_{1,i_1} \cdots \partial e_{m,i_m}}(v_1, \dots, v_m) \\ &= \sum_{i_1, \dots, i_m=1}^n \left[\sum_{\pi \in S_m} (\operatorname{sgn} \pi) v_{\pi(1),i_1} \cdots v_{\pi(m),i_m} \right] \frac{\partial^m f}{\partial e_{1,i_1} \cdots \partial e_{m,i_m}}(v_1, \dots, v_m). \end{aligned}$$

If $i_\alpha = i_\beta$ for some $\alpha \neq \beta$, then the sum inside the brackets is clearly 0. This always occurs if $m > n$, so we obtain 0 for the total sum in this case. When $m = n$, the sum inside the brackets is zero unless i_1, \dots, i_n is a permutation of $1, \dots, n$, so our total sum becomes

$$\begin{aligned} & \sum_{\rho \in S_n} \left[\sum_{\pi \in S_n} (\operatorname{sgn} \pi) v_{\pi(1),\rho(1)} \cdots v_{\pi(n),\rho(n)} \right] \frac{\partial^n f}{\partial e_{1,\rho(1)} \cdots \partial e_{n,\rho(n)}}(v_1, \dots, v_n) \\ &= \sum_{\rho \in S_n} (\operatorname{sgn} \rho) \cdot \det(v_1, \dots, v_n) \cdot \frac{\partial^n f}{\partial e_{1,\rho(1)} \cdots \partial e_{n,\rho(n)}}(v_1, \dots, v_n) \\ &= \det(v_1, \dots, v_n) \cdot \Omega f(v_1, \dots, v_n). \quad \spadesuit \end{aligned}$$

In order to make use of the Capelli identities, we introduce a partial ordering $<$ on the homogeneous polynomial functions of m vectors in \mathbb{R}^n . Let f and \bar{f} be homogeneous of degrees $(\alpha_1, \dots, \alpha_m)$ and $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$, respectively, and set $d = \alpha_1 + \cdots + \alpha_m$ and $\bar{d} = \bar{\alpha}_1 + \cdots + \bar{\alpha}_m$. Then $f < \bar{f}$ if and only if: $d < \bar{d}$; or $d = \bar{d}$ and $\alpha_m < \bar{\alpha}_m$; or $d = \bar{d}$ and $\alpha_m = \bar{\alpha}_m$ and $\alpha_{m-1} < \bar{\alpha}_{m-1}$; or \dots . This can also be expressed a little differently. Among all homogeneous polynomial functions f of fixed **total degree** $d = \alpha_1 + \cdots + \alpha_m$, we can consider $(\alpha_m, \alpha_{m-1}, \dots, \alpha_1)$ as the digits of a number to the base $d + 1$. We define the **rank** of f to be this number,

$$\operatorname{rank} f = \alpha_1 + \alpha_2(d + 1) + \alpha_3(d + 1)^2 + \cdots + \alpha_m(d + 1)^{m-1}.$$

If f and g both have total degree d , then $f < g$ if and only if $\operatorname{rank} f < \operatorname{rank} g$.

Now consider the effect on f of the operator on the left side of the Capelli identities. The main term

$$(D_{11} + m - 1) \cdots D_{mm} f$$

is (by Euler's theorem) just

$$(\alpha_1 + m - 1) \cdots a_m f = (\text{constant}) \cdot f,$$

and this constant is 0 only if f does not depend on v_m . All other terms will involve certain diagonal terms, which are all just multiplications by constants, and a term

$$D_{s_1 r_1} \cdots D_{s_\mu r_\mu} f \quad \text{where} \quad \begin{cases} r_1 < \cdots < r_\mu \\ s_i \neq r_i \\ s_1, \dots, s_\mu \text{ is a permutation} \\ \text{of } r_1, \dots, r_\mu. \end{cases}$$

In particular, $s_\mu < r_\mu$. But $D_{s_\mu r_\mu} f$ is homogeneous of degree

$$(\alpha_1, \dots, \alpha_{s_\mu} + 1, \dots, \alpha_{r_\mu} - 1, \dots, \alpha_m),$$

which means that $D_{s_\mu r_\mu} f < f$. Thus

$$D_{s_1 r_1} \cdots D_{s_\mu r_\mu} f = D_{s_1 r_1} \cdots f^* = \mathcal{P} f^*,$$

where $f^* < f$, and \mathcal{P} is a composition of polarizations; since f^* is itself a polarization of f , it is invariant under $O(n)$ or $SO(n)$ if f is. So the Cappeli identities show that

$$(A) \quad (\text{constant}) \cdot f = \begin{cases} \text{a sum of terms } \mathcal{P} f^* & m > n \\ \text{a sum of terms } \mathcal{P} f^* + \det \cdot \Omega f & m = n \end{cases}$$

where $f^* < f$ is invariant under $O(n)$ or $SO(n)$ if f is, \mathcal{P} is a composition of polarizations, and the constant is 0 only if f does not depend on v_m .

We are now ready to prove

35. THEOREM. For all m and n we have

O_n^m : Every polynomial function f of m vectors in \mathbb{R}^n which is invariant under $O(n)$ can be written as a polynomial in the inner product functions ι_{rs} .

SO_n^m : Every polynomial function f of m vectors in \mathbb{R}^n which is invariant under $SO(n)$ can be written as a polynomial in the functions ι_{rs} and the determinant functions $\det_{r_1 \dots r_n}$.

PROOF. Notice that a function of m vectors can always be thought of as a function of a larger number of vectors; so $O_n^m \implies O_n^{m'}$ and $SO_n^m \implies SO_n^{m'}$ automatically for $m' < m$. Note also that the determinants in SO_n^m are zero unless $m \geq n$.

The proof of O_n^m and SO_n^m proceeds in two parts.

36. LEMMA. If O_n^n [respectively SO_n^n] holds, then O_n^m [respectively SO_n^m] holds for all $m > n$.

PROOF. The proof for SO will be almost exactly the same as for O, so we give only the latter. Actually, we give the proof only for O_n^{n+1} , as it will then be clear how to proceed by induction. We consider invariant homogeneous polynomial functions f of $n+1$ vectors in \mathbb{R}^n , of fixed total degree d . We will prove that they can be represented in the desired form by complete induction on their rank. If $\text{rank } f < (d+1)^n$, so that the degree α_n of f in v_{n+1} is 0, then f does not involve v_{n+1} , so the result follows from the hypothesis that O_n^n holds. Let $r_0 \geq (d+1)^n$. Assuming that all f of total degree d and rank $< r_0$ can be expressed in the desired form, we will show that all f of total degree d and rank r_0 can also be so expressed. The constant in equation (A) is $\neq 0$ for our f , so f is the sum of terms $\mathcal{P}f^*$, where $f^* < f$ is invariant under $O(n)$ and \mathcal{P} is a composition of polarizations. Since f^* is a single polarization applied to f , the total degree of f^* equals the total degree of f . Since $f^* < f$, the inductive assumption says that each f^* can be written

$$f^* = F^* \circ (\{t_{rs}\})$$

for some polynomial F^* . This implies that

$$\mathcal{P}f^* = \mathcal{F}^* \circ (\{\mathcal{P}t_{rs}\})$$

for some polynomial \mathcal{F}^* . Since each $\mathcal{P}t_{rs}$ is a sum of t_{rs} 's, it follows that each $\mathcal{P}f^*$ is a polynomial in the t_{rs} , and thus f is. **Q.E.D.**

We still have to show that O_n^n and SO_n^n hold. In the case of SO_n^n there is a single determinant $\det(v_{ri})$ involved. If $A = (v_{ri})$, then

$$(A \cdot A^t)_{rs} = \sum_k v_{rk} v_{sk} = \langle v_r, v_s \rangle.$$

Hence

$$[\det(v_{ri})]^2 = \det A \cdot A^t = \det(\langle v_r, v_s \rangle)$$

is a polynomial in the t_{rs} . Thus we need only linear terms in \det .

37. LEMMA. O_n^n holds for all n . Moreover, SO_n^n holds in the strengthened form

SO_n^n : Every polynomial function of n vectors in \mathbb{R}^n which is invariant under $SO(n)$ can be written as $g + (\det) \cdot h$, where g and h are polynomials in the inner product functions t_{rs} .

PROOF. We use induction on n . Consider first a polynomial function $f: \mathbb{R} \rightarrow \mathbb{R}$. If f is invariant under $O(1)$, then $f(x) = f(-x)$. So $f(x)$ involves only even powers of x . Hence $f(x) = F(x^2) = F(\langle x, x \rangle)$ for some polynomial F . Moreover, *any* polynomial function $f: \mathbb{R} \rightarrow \mathbb{R}$ can be written

$$f(x) = g(x) + xh(x) = g(x) + (\det x) \cdot h(x),$$

where g and h involve only even powers.

To carry out the induction step,

$$(*) \quad \{O_{n-1}^{n-1} \text{ and } SO_{n-1}^{n-1}\} \implies \{O_n^n \text{ and } SO_n^n\},$$

we first show that

$$O_{n-1}^{n-1} \implies SO_n^{n-1} \quad (\implies O_n^{n-1}, \text{ since no determinants are involved}).$$

So consider a polynomial function f of $n-1$ vectors in \mathbb{R}^n . Define a polynomial function \bar{f} of $n-1$ vectors in \mathbb{R}^{n-1} by

$$\bar{f}(w_1, \dots, w_{n-1}) = f(\bar{w}_1, \dots, \bar{w}_{n-1}),$$

where

$$\bar{w}_r = (w_{r1}, \dots, w_{r,n-1}, 0).$$

If f is invariant under $SO(n)$, then \bar{f} is actually invariant under $O(n-1)$. So by hypothesis, there is a polynomial \bar{F} with

$$\bar{f}(w_1, \dots, w_{n-1}) = \bar{F}(\{\langle w_r, w_s \rangle\}).$$

Now given $v_1, \dots, v_{n-1} \in \mathbb{R}^n$, choose $A \in SO(n)$ so that all $A \cdot v_r$ lie in $\mathbb{R}^{n-1} \times \{0\}$, and hence $A \cdot v_r = \bar{w}_r$ for some $w_r \in \mathbb{R}^{n-1}$. Then

$$\begin{aligned} f(v_1, \dots, v_{n-1}) &= f(A \cdot v_1, \dots, A \cdot v_{n-1}) = f(\bar{w}_1, \dots, \bar{w}_{n-1}) \\ &= \bar{f}(w_1, \dots, w_n) = \bar{F}(\{\langle w_r, w_s \rangle\}) \\ &= \bar{F}(\{\langle v_r, v_s \rangle\}). \end{aligned}$$

This completes the proof that $O_{n-1}^{n-1} \implies SO_n^{n-1}$.

Now for the proof of (*). This proof will also be by induction, using the same general scheme as in the previous lemma, but there will be a slight complication, for we will actually be using a double induction, first on the total degree of f , and then within each total degree on the rank. In addition, the statements O_n^n and SO_n^n will have to be proved jointly in the induction. Thus, for a fixed total

degree and rank, we will show that all f of this degree and rank which are invariant under $O(n)$ have the desired form, and also that all f of this degree and rank which are invariant under $SO(n)$ have the desired form, assuming that the same two statements hold for all f of lower degree, or of the same degree and lower rank. There is certainly no problem beginning the induction with degree 0; moreover, within any particular degree, the polynomials of sufficiently low rank will not involve v_n , so we will be back to the cases SO_n^{n-1} and O_n^{n-1} which we have already proved.

Now consider a particular invariant f . We use equation (A), in the case $m = n$, to see that f is a sum of terms $\mathcal{P}f^*$ plus a constant times $\det \cdot \Omega f$. The sum of the terms $\mathcal{P}f^*$ can be written as $F \circ (\{\iota_{rs}\})$, as before, and we just have to worry about $\det \cdot \Omega f$. First suppose that f is invariant under $O(n)$. Then $\Omega f \prec f$ is invariant under $SO(n)$, so by the induction hypothesis we can write

$$\Omega f = g + (\det) \cdot h,$$

where $g, h \prec f$ are invariant under $O(n)$, and thus by the induction hypothesis expressible as polynomials in the ι_{rs} . So we have

$$f = F \circ (\{\iota_{rs}\}) + (\text{constant}) \cdot \det \cdot [G \circ (\{\iota_{rs}\}) + \det \cdot H \circ (\{\iota_{rs}\})].$$

Since f, ι_{rs} , and \det^2 are invariant under $O(n)$, the term $\det \cdot G \circ (\{\iota_{rs}\})$ must also be invariant under $O(n)$, which is possible only if $G = 0$. Thus f is a polynomial in the ι_{rs} .

If f is assumed invariant under $SO(n)$, then everything remains the same, except that G need not be zero. ♦

8. AN EASIER INVARIANCE PROBLEM

For an $n \times n$ matrix M , we define $f_1(M), \dots, f_n(M)$ by

$$\det(I + \lambda M) = 1 + \lambda f_1(M) + \dots + \lambda^n f_n(M).$$

It will also be convenient to set $f_0(M) = 1$. Then the f_i are polynomial functions of the entries of M which are invariant under the adjoint action of $GL(n, \mathbb{R})$,

$$f_i(\text{Ad}(A)M) = f_i(AMA^{-1}) = f_i(M) \quad \text{for } A \in GL(n, \mathbb{R}).$$

If M has eigenvalues $\lambda_1, \dots, \lambda_n$, and σ_i denotes the i^{th} elementary symmetric polynomial, then*

$$f_i(M) = \sigma_i(\lambda_1, \dots, \lambda_n).$$

According to Problem I.7-15, every polynomial function on the $n \times n$ matrices $\mathfrak{gl}(n, \mathbb{R})$ which is invariant under the adjoint action of $\text{GL}(n, \mathbb{R})$ is a polynomial in the f_i (notice that we are now considering functions of a *single matrix*, rather than functions of *many vectors*). Now we want to find out which polynomial functions on $\mathfrak{o}(n)$ are invariant under the adjoint action of $\text{O}(n)$, or of $\text{SO}(n)$. The line of argument will be essentially that used in Problem I.7-15, except that in some ways it will be even easier, since we have an especially simple “canonical form” for elements of $\mathfrak{o}(n)$, which greatly strengthens Corollary 11:

38. PROPOSITION. For every $A \in \mathfrak{o}(n)$ there is a matrix $B \in \text{O}(n)$ such that BAB^{-1} equals either

$$\begin{pmatrix} 0 & \lambda_1 & & & 0 \\ -\lambda_1 & 0 & & & \\ & & \ddots & & \\ 0 & & & 0 & \lambda_m \\ & & & -\lambda_m & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & \lambda_1 & & & 0 \\ -\lambda_1 & 0 & & & \\ & & \ddots & & \\ 0 & & & 0 & \lambda_m \\ & & & -\lambda_m & 0 \\ & & & & & 0 \end{pmatrix}$$

(some of the λ 's may also be 0).

PROOF. If $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the linear transformation determined by A , then the skew-symmetry of A means that

$$(1) \quad \langle Tv, w \rangle = -\langle v, Tw \rangle \quad \text{for } v, w \in \mathbb{R}^n.$$

[Conversely, if this relation holds for $T: (V, \langle \cdot, \cdot \rangle) \rightarrow (V, \langle \cdot, \cdot \rangle)$, then the matrix of T with respect to an orthonormal basis is skew-symmetric, and we have

$$(2) \quad \det T = 0 \quad \text{if } \dim V \text{ is odd}$$

as a particular consequence.] Equation (1) implies that $(\text{image } T)^\perp = \ker T$. Consequently, T must be one-one on $\text{image } T$. Therefore $\text{rank } T^2 = \text{rank } T$.

*The λ_i are the roots of the characteristic polynomial $\chi(\lambda) = \det(\lambda I - M)$. Recall that the $\sigma_i = \sigma_i(\lambda_1, \dots, \lambda_n)$ satisfy $\lambda^n - \sigma_1 \lambda^{n-1} + \dots = \prod_i (\lambda - \lambda_i) = \chi(\lambda)$. So $\lambda^n + \sigma_1 \lambda^{n-1} + \dots = (-1)^n \chi(-\lambda) = (-1)^n \det(-\lambda I - M) = \det(\lambda I + M)$. Hence, $\det(I + \lambda M) = \lambda^n \det(I/\lambda + M) = \lambda^n [(1/\lambda)^n + \sigma_1 (1/\lambda)^{n-1} + \dots] = 1 + \sigma_1 \lambda + \dots$.

Moreover, this rank is even, by Corollary 11. [Alternate proof: The map $T|(\text{image } T): \text{image } T \rightarrow \text{image } T$ is one-one, so its determinant must be non-zero. Applying (2), we see that $\dim \text{image } T$ is even.]

Now A^2 is symmetric, so there is an orthonormal basis v_1, \dots, v_n of eigenvectors of T^2 , with eigenvalues μ_1, \dots, μ_n . Note that

$$\mu_j = \langle v_j, T^2 v_j \rangle = -\langle T v_j, T v_j \rangle \leq 0.$$

By renumbering, we can assume that $\mu_1, \dots, \mu_{2m} < 0$, and that the remaining μ 's equal 0. Define a new orthonormal basis w_1, \dots, w_n by

$$\begin{aligned} w_1 &= v_1, & w_2 &= \frac{1}{\sqrt{-\lambda_1}} T(v_1) \\ &\vdots & & \\ w_{2m-1} &= v_m, & w_{2m} &= \frac{1}{\sqrt{-\lambda_m}} T(v_m) \\ w_j &= v_j & j &> 2m. \end{aligned}$$

The matrix of T has the desired form with respect to the basis w_1, \dots, w_n . ♦

Notice that for $M \in \mathfrak{o}(n)$ we have

$$(I + \lambda M)^t = I - \lambda M \implies \det(I + \lambda M) = \det(I - \lambda M),$$

so

$$1 + \lambda f_1(M) + \dots + \lambda^n f_n(M) = 1 - \lambda f_1(M) + \dots + (-1)^n \lambda^n f_n(M),$$

and hence $f_i(M) = 0$ for odd i . We will also need to use the following formula, whose verification is left to the reader:

$$\sigma_{2k}(\lambda_1, -\lambda_1, \lambda_2, -\lambda_2, \dots, \lambda_m, -\lambda_m) = \sigma_k(-\lambda_1^2, \dots, -\lambda_m^2).$$

39. THEOREM. Let $n = 2m$ or $n = 2m + 1$. Then every polynomial function f on $\mathfrak{o}(n)$ which is invariant under the adjoint action of $O(n)$ is a polynomial in f_2, \dots, f_{2m} .

PROOF. For $\lambda_1, \dots, \lambda_m \in \mathbb{R}$, let $[\lambda_1, \dots, \lambda_m]$ be

$$\begin{pmatrix} 0 & \lambda_1 & & 0 \\ -\lambda_1 & 0 & & \\ & & \ddots & \\ 0 & & & 0 & \lambda_m \\ & & & -\lambda_m & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & \lambda_1 & & 0 \\ -\lambda_1 & 0 & & \\ & & \ddots & \\ 0 & & & 0 & \lambda_m \\ & & & -\lambda_m & 0 \\ & & & & 0 \end{pmatrix},$$

depending on whether $n = 2m$ or $n = 2m + 1$. Notice that the eigenvalues of $[\lambda_1, \dots, \lambda_m]$ are $i\lambda_1, -i\lambda_1, \dots, i\lambda_m, -i\lambda_m$ [and 0, if $n = 2m + 1$], so

$$\begin{aligned} f_{2k}([\lambda_1, \dots, \lambda_m]) &= \sigma_{2k}(i\lambda_1, -i\lambda_1, \dots, i\lambda_m, -i\lambda_m) \\ &= \sigma_k(\lambda_1^2, \dots, \lambda_m^2). \end{aligned}$$

Define

$$g(\lambda_1, \dots, \lambda_m) = f([\lambda_1, \dots, \lambda_m]).$$

Then g is a polynomial function of $\lambda_1, \dots, \lambda_m$. Notice that for the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & & 0 \\ & 0 & 1 & & \\ 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & & & 1 & \\ 0 & & & & \ddots & \\ & & & & & 1 \end{pmatrix} \in O(n)$$

we have

$$A \cdot [\lambda_1, \dots, \lambda_m] \cdot A^{-1} = [\lambda_2, \lambda_1, \dots, \lambda_m].$$

Similarly, we can interchange any two λ 's by some $A \in O(n)$. Thus g is symmetric in the λ 's. Notice, moreover, that for the matrix

$$B = \begin{pmatrix} 0 & 1 & & 0 \\ 1 & 0 & & \\ & & 1 & \\ 0 & & & \ddots & \\ & & & & 1 \end{pmatrix} \in O(n)$$

we have

$$B \cdot [\lambda_1, \dots, \lambda_m] B^{-1} = [-\lambda_1, \lambda_2, \dots, \lambda_m].$$

Thus

$$g(\lambda_1, \dots, \lambda_m) = g(-\lambda_1, \lambda_2, \dots, \lambda_m).$$

This shows that the polynomial g does not have any terms involving λ_1 to an odd power. The same result clearly holds for all λ 's, so we can write

$$g(\lambda_1, \dots, \lambda_m) = h(\lambda_1^2, \dots, \lambda_m^2)$$

for some polynomial h . Clearly h is symmetric in its arguments, so there is a polynomial p with

$$g(\lambda_1, \dots, \lambda_m) = p(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_m(\lambda_1^2, \dots, \lambda_m^2)).$$

Thus we have

$$f([\lambda_1, \dots, \lambda_m]) = p(f_2([\lambda_1, \dots, \lambda_m]), \dots, f_{2m}([\lambda_1, \dots, \lambda_m])).$$

Now for any $M \in \mathfrak{o}(n)$ there is, by Proposition 38, some $A \in O(n)$ such that $A^{-1}MA = [\lambda_1, \dots, \lambda_m]$ for some $\lambda_1, \dots, \lambda_m$. Since f, f_2, \dots, f_{2m} are invariant under the adjoint action of $O(n)$, the above equation yields

$$f(M) = p(f_2(M), \dots, f_{2m}(M)). \quad \blacklozenge$$

Note that the polynomial functions f_2, \dots, f_{2m} are algebraically independent on $\mathfrak{o}(n)$ —if

$$p(f_2(M), \dots, f_{2m}(M)) = 0$$

for all $M \in \mathfrak{o}(n)$, then $p = 0$. Indeed, this equation implies that

$$p(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_m(\lambda_1^2, \dots, \lambda_m^2)) = 0$$

for all $\lambda_1, \dots, \lambda_m$, and thus that the polynomial

$$p(\sigma_1(x_1, \dots, x_m), \dots, \sigma_m(x_1, \dots, x_m))$$

is zero whenever x_1, \dots, x_m take on positive values; but this implies that this polynomial in x_1, \dots, x_m is identically zero, and hence that $p = 0$, as we remarked on page 318.

With slight modifications of our previous argument we obtain

40. THEOREM. (1) If $n = 2m + 1$, then every polynomial function f on $\mathfrak{o}(n)$ which is invariant under the adjoint action of $\mathrm{SO}(n)$ is a polynomial in f_2, \dots, f_{2m} .

(2) If $n = 2m$, then every polynomial function f on $\mathfrak{o}(n)$ which is invariant under the adjoint action of $\mathrm{SO}(n)$ is a polynomial in f_2, \dots, f_{2m-2} and the Pfaffian Pf .

PROOF. (1) Notice that the matrix A in the proof of Theorem 39 is actually in $\mathrm{SO}(n)$. If n is odd, then $[\lambda_1, \dots, \lambda_m]$ has a zero in the $(n, n)^{\mathrm{th}}$ place, so for the matrix B in the proof of Theorem 39 we can just as well replace the 1 in the $(n, n)^{\mathrm{th}}$ place by -1 ; this new B is in $\mathrm{SO}(n)$. Now the proof of Theorem 39 goes through as before.

(2) The matrix A is still in $\mathrm{SO}(n)$. We cannot arrange for B to be in $\mathrm{SO}(n)$, but for the matrix

$$C = \begin{pmatrix} 0 & 1 & & & & \\ 1 & 0 & & & & \\ & & 0 & 1 & & \\ & & 1 & 0 & & \\ & & & & 1 & \\ & & & & & \ddots \\ & & & & & & 1 \end{pmatrix} \in \mathrm{SO}(n),$$

we have

$$C \cdot [\lambda_1, \dots, \lambda_m] C^{-1} = [-\lambda_1, -\lambda_2, \lambda_3, \dots, \lambda_m].$$

Similarly, we can send any pair of λ 's to their negatives. So the symmetric function g in the proof of Theorem 39 has the property that each monomial appearing in it is either of even degree in all λ 's or else of odd degree in all λ 's. So g can be written

$$g(\lambda_1, \dots, \lambda_m) = h_1(\lambda_1^2, \dots, \lambda_m^2) + (\lambda_1 \cdots \lambda_m) h_2(\lambda_1^2, \dots, \lambda_m^2).$$

Since g is symmetric, the term $h_1(\lambda_1^2, \dots, \lambda_m^2)$ [= the sum of the monomials of g which are of even degree in all λ 's] must be symmetric in $\lambda_1, \dots, \lambda_m$. So h_1 is symmetric in its arguments. Thus h_2 is also symmetric in its arguments. So we can write

$$\begin{aligned} g(\lambda_1, \dots, \lambda_m) &= p_1(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_m(\lambda_1^2, \dots, \lambda_m^2)) \\ &\quad + (\lambda_1 \cdots \lambda_m) \cdot p_2(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_m(\lambda_1^2, \dots, \lambda_m^2)). \end{aligned}$$

Thus we have

$$f([\lambda_1, \dots, \lambda_m]) = p_1(f_2([\lambda_1, \dots, \lambda_m], \dots, f_{2m}([\lambda_1, \dots, \lambda_m])) \\ + \text{Pf}([\lambda_1, \dots, \lambda_m]) \cdot p_2(f_2([\lambda_1, \dots, \lambda_m]), \dots, f_{2m}([\lambda_1, \dots, \lambda_m]))).$$

It follows, as before, that

$$f(M) = p_1(f_2(M), \dots, f_{2m}(M)) \\ + \text{Pf}(M) \cdot p_2(f_2(M), \dots, f_{2m}(M))$$

for all $M \in \mathfrak{o}(n)$. We can dispense with f_{2m} , since

$$f_{2m}(M) = \det M = \{\text{Pf}(M)\}^2. \spadesuit$$

We have already observed that the polynomials f_2, \dots, f_{2m} are algebraically independent on $\mathfrak{o}(n)$. For $n = 2m$, the polynomials $f_2, \dots, f_{2m-2}, \text{Pf}$ are algebraically independent on $\mathfrak{o}(n)$. For suppose that

$$p(f_2(M), \dots, f_{2m-2}(M), \text{Pf}(M)) = 0 \quad \text{for all } M \in \mathfrak{o}(n).$$

Then for all $\lambda_1, \dots, \lambda_m$ we have

$$0 = p(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_{m-1}(\lambda_1^2, \dots, \lambda_m^2), \lambda_1 \cdots \lambda_m) \\ = p_1(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_{m-1}(\lambda_1^2, \dots, \lambda_m^2), \sigma_m(\lambda_1^2, \dots, \lambda_m^2)) + \\ (\lambda_1 \cdots \lambda_m) p_2(\sigma_1(\lambda_1^2, \dots, \lambda_m^2), \dots, \sigma_{m-1}(\lambda_1^2, \dots, \lambda_m^2), \sigma_m(\lambda_1^2, \dots, \lambda_m^2)),$$

for certain polynomials p_1 and p_2 . This polynomial in $\lambda_1, \dots, \lambda_m$ can be zero only if the two summands, representing the terms with all λ 's of even degree and the terms with all λ 's of odd degree, respectively, are each zero. Then as before we conclude that $p_1 = p_2 = 0$.

There is one further simple property of the functions f_k . In Corollary 13 we gave a formula for $\text{Pf}(A \oplus B)$, where

$$A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

It is easy to find a formula for $f_k(A \oplus B)$, for all $A \in \mathfrak{gl}(r, \mathbb{R})$ and $B \in \mathfrak{gl}(s, \mathbb{R})$. For we have

$$\det(I_{r+s} + \lambda(A \oplus B)) \\ = \det((I_r + \lambda A) \oplus (I_s + \lambda B)) \\ = \det(I_r + \lambda A) \cdot \det(I_s + \lambda B) \\ = (1 + \lambda f_1(A) + \cdots + \lambda^r f_r(A)) \cdot (1 + \lambda f_1(B) + \cdots + \lambda^s f_s(B)).$$

So

$$f_k(A \oplus B) = \text{coefficient of } \lambda^k = \sum_{l=0}^k f_l(A) \cdot f_{k-l}(B).$$

When A and B are skew-symmetric, we have

$$f_{2k}(A \oplus B) = \sum_{l=1}^k f_{2l}(A) \cdot f_{2k-2l}(B).$$

9. THE COHOMOLOGY OF THE ORIENTED GRASSMANNIANS

We are now ready to compute part of the cohomology of

$$\tilde{G}_n(\mathbb{R}^N) = \text{SO}(N)/\text{SO}(n) \times \text{SO}(N-n) = \text{SO}(N)/H.$$

The Lie algebra $\mathfrak{o}(N)$ has as a basis the matrices

$$X_{\alpha}^{\beta} = \begin{matrix} & \alpha & \beta \\ \alpha & & \\ \beta & \begin{pmatrix} & 1 \\ -1 & \end{pmatrix} \end{matrix} \quad 1 \leq \alpha < \beta \leq N$$

which have non-zero entries only in the (α, β) and (β, α) positions. [We adopt the convention that the indices α, β range from 1 to N , while the indices i, j run from 1 to n , and r, s range from $n+1$ to N .] Let $\{\phi_{\alpha}^{\beta}\}$ be the dual basis to the $\{X_{\alpha}^{\beta}\}$. The Lie algebra \mathfrak{h} consists of matrices

$$\begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix}, \quad \begin{matrix} L_1 \in \mathfrak{o}(n) \\ L_2 \in \mathfrak{o}(N-n). \end{matrix}$$

The orthogonal complement \mathfrak{h}^{\perp} , with respect to the bi-invariant metric on page 308, consists of all matrices

$$\begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix}, \quad P \text{ an } n \times (N-n) \text{ matrix,}$$

and has as basis the X_i^r for $1 \leq i \leq n$ and $n+1 \leq r \leq N$; so the corresponding ϕ_i^r are a basis for $(\mathfrak{h}^{\perp})^*$ [more precisely, the restrictions of the ϕ_i^r to \mathfrak{h}^{\perp} are a basis].

The adjoint action of $\mathrm{SO}(n) \times \mathrm{SO}(N - n)$ on \mathfrak{h}^\perp is easily computed to be

$$\begin{aligned} \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}^{-1} &= \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \begin{pmatrix} A^t & 0 \\ 0 & B^t \end{pmatrix} \\ &= \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} 0 & PB^t \\ -P^t A^t & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & APB^t \\ -BP^t A^t & 0 \end{pmatrix}. \end{aligned}$$

We want to know which elements of $\Omega^k(\mathfrak{h}^\perp)$ are invariant under the induced adjoint action of $\mathrm{SO}(n) \times \mathrm{SO}(N - n)$. We will split this question up into two parts, by considering invariance under the adjoint action of the two subgroups

$$\begin{aligned} \mathrm{SO}(n) \times \{I\} &= \left\{ \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} : A \in \mathrm{SO}(n) \right\} \\ \{I\} \times \mathrm{SO}(N - n) &= \left\{ \begin{pmatrix} I & 0 \\ 0 & B \end{pmatrix} : B \in \mathrm{SO}(N - n) \right\}. \end{aligned}$$

We consider first the adjoint action of $\{I\} \times \mathrm{SO}(N - n)$, given by

$$\begin{pmatrix} I & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & B \end{pmatrix}^{-1} = \begin{pmatrix} 0 & PB^t \\ -BP^t & 0 \end{pmatrix}.$$

If we regard \mathfrak{h}^\perp as the n -fold product $\mathbb{R}^{N-n} \times \cdots \times \mathbb{R}^{N-n}$ by identifying $\begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \in \mathfrak{h}^\perp$ with the n -tuple of the rows of P , then this action is the usual action of $\mathrm{SO}(N - n)$ on each factor. Since a form $\eta \in \Omega^k(\mathfrak{h}^\perp)$ can be regarded as a polynomial function on the nk -fold product $\mathbb{R}^{N-n} \times \cdots \times \mathbb{R}^{N-n}$, Theorem 35 shows that η is invariant under the adjoint action of $\{I\} \times \mathrm{SO}(N - n)$ if and only if it is a polynomial in the inner products and determinants of the vectors involved. We have to figure out just what this means when η is a k -form, and express η in terms of the forms ϕ_i^r . *From now on we assume that $k < N - n$.*

Consider first the case where η is a 1-form, and thus a function $\eta: \mathfrak{h}^\perp \rightarrow \mathbb{R}$. If η is invariant under $\{I\} \times \mathrm{SO}(N - n)$, and $M = \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix}$, then $\eta(M)$ can be written as a polynomial in the inner products of rows of P and in the $(N - n) \times (N - n)$ subdeterminants of P . Thus $\eta(M)$ is a polynomial in

$$\sum_{r=n+1}^N \phi_{i_1}^r(M) \cdot \phi_{i_2}^r(M) \quad 1 \leq i_1, i_2 \leq n$$

and

the determinants of the matrices formed
by picking $(N - n)$ rows of P .

Multiplying M by $\alpha \in \mathbb{R}$ multiplies the first terms by α^2 and the determinants by α^{N-n} . So $\eta(M)$ cannot be linear in M unless it is zero. Thus $\Omega^1(\mathfrak{h}^\perp)^H = 0$.

Now consider a 2-form $\eta \in \Omega^2(\mathfrak{h}^\perp)$. If η is invariant under $\{I\} \times \text{SO}(N - n)$, then $\eta(M_1, M_2)$ can be written as a polynomial in

$$\sum_r \phi_{i_1}^r(M_1) \cdot \phi_{i_2}^r(M_1), \quad \sum_r \phi_{i_1}^r(M_2) \cdot \phi_{i_2}^r(M_2), \quad \sum_r \phi_{i_1}^r(M_1) \cdot \phi_{i_2}^r(M_2)$$

and

the determinants of the matrices formed by picking n_1 rows
of P_1 and n_2 rows of P_2 , with $n_1 + n_2 = (N - n)$.

Multiplying M_1 [or M_2] by α multiplies these determinants by α^{n_1} [or α^{n_2}]. But either $n_1 > 1$ or $n_2 > 1$, since we are assuming that $2 = k < N - n$. Consequently, since η is multilinear, the determinants cannot be involved. Moreover, of the remaining terms, only those of the third kind can be involved. So

$$\eta = \text{a linear combination of the } \sum_r \phi_{i_1}^r \otimes \phi_{i_2}^r.$$

Since η is a 2-form, we have

$$\eta = \text{Alt } \eta = \text{a linear combination of the } \sum_r \phi_{i_1}^r \wedge \phi_{i_2}^r.$$

Thus $\Omega^2(\mathfrak{h}^\perp)^H$ can contain only linear combinations of the 2-forms

$$\zeta_{i_1 i_2} = \sum_r \phi_{i_1}^r \wedge \phi_{i_2}^r \quad 1 \leq i_1 < i_2 \leq n.$$

For a 3-form $\eta \in \Omega^3(\mathfrak{h}^\perp)$ to be invariant under $\{I\} \times \text{SO}(N - n)$, it must be possible to write $\eta(M_1, M_2, M_3)$ as a polynomial in

$$\sum_r \phi_{i_1}^r(M_{j_1}) \cdot \phi_{i_2}^r(M_{j_2}) \quad j_1, j_2 = 1, 2, 3$$

and

determinants formed from rows of P_1, P_2, P_3 .

As before, the determinants cannot be involved, since we assume $3 = k < N - n$. Then it is easy to see that no non-zero polynomial in the other terms can be multilinear in (M_1, M_2, M_3) . So $\Omega^3(\mathfrak{h}^\perp)^H = 0$.

If $\eta \in \Omega^4(\mathfrak{h}^\perp)$ is invariant under $\{I\} \times \mathrm{SO}(N - n)$, then $\eta(M_1, \dots, M_4)$ can be written as a polynomial in the

$$\sum_r \phi_{i_1}^r(M_{j_1}) \cdot \phi_{i_2}^r(M_{j_2}) \quad j_1, j_2 = 1, \dots, 4$$

(determinants are ruled out as before). Since η is multilinear, it is easy to see that the only monomials which can appear are

$$\left\{ \sum_r \phi_{i_1}^r(M_{j_1}) \cdot \phi_{i_2}^r(M_{j_2}) \right\} \cdot \left\{ \sum_r \phi_{i_3}^r(M_{j_3}) \cdot \phi_{i_4}^r(M_{j_4}) \right\},$$

where j_1, \dots, j_4 are distinct. This term can be written

$$\left(\sum_r \phi_{i_1}^r \otimes \phi_{i_2}^r \right) \otimes \left(\sum_r \phi_{i_3}^r \otimes \phi_{i_4}^r \right) (M_{\pi(1)}, M_{\pi(2)}, M_{\pi(3)}, M_{\pi(4)})$$

for some permutation $\pi \in S_4$. Since η is alternating, we find that η is a linear combination of terms $\zeta_{i_1 i_2} \wedge \zeta_{i_3 i_4}$.

In general, we clearly have:

If $k < N - n$ is odd, then $\Omega^k(\mathfrak{h}^\perp)^H = 0$.

If $k < N - n$ is even, then all elements of $\Omega^k(\mathfrak{h}^\perp)^H$ can be written as linear combinations of the forms

$$\zeta_{i_1 i_2} \wedge \dots \wedge \zeta_{i_{k-1} i_k}.$$

To determine $\Omega^k(\mathfrak{h}^\perp)^H$ completely for all even $k < N - n$, we still have to consider invariance under $\mathrm{SO}(n) \times \{I\}$. But we already see, from Theorem 32, that

(A) If $k < N - n$ is odd, then $H^k(\tilde{G}_n(\mathbb{R}^N)) = 0$.

Moreover, the maps

$$\Omega^{k-1}(\mathfrak{h}^\perp)^H \xrightarrow{d} \Omega^k(\mathfrak{h}^\perp)^H \xrightarrow{d} \Omega^{k+1}(\mathfrak{h}^\perp)^H$$

are zero for even $k < N - n - 1$, since the vector spaces on the ends are 0. Consequently,

$$H^k(\tilde{G}_n(\mathbb{R}^N)) = \frac{\ker d: \Omega^k(\mathfrak{h}^\perp)^H}{d(\Omega^{k+1}(\mathfrak{h}^\perp)^H)} = \Omega^k(\mathfrak{h}^\perp)^H / 0 = \Omega^k(\mathfrak{h}^\perp)^H.$$

Actually, this result holds for *all* k , but we need a different argument. Notice that something special happens when we take the bracket of two elements of \mathfrak{h}^\perp . We have

$$\begin{aligned} \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \begin{pmatrix} 0 & Q \\ -Q^t & 0 \end{pmatrix} &= \begin{pmatrix} 0 & Q \\ -Q^t & 0 \end{pmatrix} \begin{pmatrix} 0 & P \\ -P^t & 0 \end{pmatrix} \\ &= \begin{pmatrix} -PQ^t + QP^t & 0 \\ 0 & -P^tQ + QP^t \end{pmatrix}, \end{aligned}$$

which is in \mathfrak{h} . Consulting the statement of Theorem 32 we see that our map

$$d: \Omega^k(\mathfrak{h}^\perp)^H \rightarrow \Omega^{k+1}(\mathfrak{h}^\perp)^H$$

is *always* 0. Thus

$$(B) \quad \text{For all } k \text{ we have } H^k(\tilde{G}_n(\mathbb{R}^N)) = \Omega^k(\mathfrak{h}^\perp)^H.$$

We now have to investigate linear combinations of the forms $\zeta_{i_1 i_2} \wedge \cdots \wedge \zeta_{i_{k-1} i_k}$. Such combinations can be described in terms of polynomial functions $f: \mathfrak{o}(n) \rightarrow \mathbb{R}$ which are homogeneous of degree $k/2$: if f is a sum of monomials

$$c \cdot \phi_{i_2}^{i_1} \cdot \phi_{i_4}^{i_3} \cdots \phi_{i_k}^{i_{k-1}} \quad 1 \leq i_{2\alpha-1} < i_{2\alpha} \leq n,$$

then $f(\zeta) \in \Omega^k(\mathfrak{h}^\perp)$ will denote the k -form which is the sum of the corresponding terms

$$c \cdot \zeta_{i_1 i_2} \wedge \zeta_{i_3 i_4} \wedge \cdots \wedge \zeta_{i_{k-1} i_k}$$

(since the ζ_{ij} are 2-forms, the \wedge products commute, so the order of the factors ϕ_j^i is irrelevant). Clearly every linear combination of the forms $\zeta_{i_1 i_2} \wedge \cdots \wedge \zeta_{i_{k-1} i_k}$ is $f(\zeta)$ for some $f: \mathfrak{o}(n) \rightarrow \mathbb{R}$. Since the ϕ_i^j are linearly independent, this f is unique for $k < N-n$. Moreover, for homogeneous polynomials $f, g: \mathfrak{o}(n) \rightarrow \mathbb{R}$ we have

$$(1) \quad (fg)(\zeta) = f(\zeta) \wedge g(\zeta).$$

We want to find out how $\text{SO}(n) \times \{I\}$ operates on $f(\zeta)$. Take first the special case $f = \phi_j^i$, so that $f(\zeta) = \phi_j^i(\zeta) = \zeta_{ij}$. For

$$\begin{aligned} \tilde{A} &= \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \in \text{SO}(n) \times \{I\} \\ M_1 &= \begin{pmatrix} 0 & P_1 \\ -P_1^t & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & P_2 \\ -P_2^t & 0 \end{pmatrix} \in \mathfrak{h}^\perp \end{aligned}$$

we have

$$\begin{aligned}
& [\text{Ad}(\tilde{A})^* \zeta_{ij}](M_1, M_2) \\
&= \sum_r \phi_i^r \wedge \phi_j^r (\text{Ad}(\tilde{A}) M_1, \text{Ad}(\tilde{A}) M_2) \\
&= \sum_r \left\{ \phi_i^r \begin{pmatrix} 0 & AP_1 \\ -P_1^t A^t & 0 \end{pmatrix} \cdot \phi_j^r \begin{pmatrix} 0 & AP_2 \\ -P_2^t A^t & 0 \end{pmatrix} - \cdots \right\} \\
&= \sum_r \{ (AP_1)_{ir} (AP_2)_{jr} - (AP_2)_{ir} (AP_1)_{jr} \} \\
&= \sum_r \sum_{\mu, v=1}^n \{ A_{i\mu} (P_1)_{\mu r} A_{jv} (P_2)_{vr} - A_{i\mu} (P_2)_{\mu r} A_{jv} (P_1)_{vr} \} \\
&= \sum_{\mu, v=1}^n A_{i\mu} A_{jv} \left[\sum_r (P_1)_{\mu r} (P_2)_{vr} - (P_2)_{\mu r} (P_1)_{vr} \right],
\end{aligned}$$

and hence

$$\text{Ad}(\tilde{A})^* \zeta_{ij} = \sum_{\mu, v=1}^n A_{i\mu} A_{jv} \zeta_{\mu v},$$

or

$$(2) \quad \text{Ad}(\tilde{A})^* (\phi_j^i(\zeta)) = \left(\sum_{\mu, v=1}^n A_{i\mu} A_{jv} \phi_v^\mu \right) (\zeta).$$

On the other hand, for a matrix $L \in \mathfrak{o}(n)$ we have

$$\begin{aligned}
(3) \quad \phi_j^i(\text{Ad}(A)L) &= \phi_j^i(ALA^t) = \sum_{\mu, v=1}^n A_{i\mu} A_{jv} L_v^\mu \\
&\implies \phi_j^i \circ \text{Ad}(A) = \sum_{\mu, v=1}^n A_{i\mu} A_{jv} \phi_v^\mu.
\end{aligned}$$

Comparing (2) and (3), we see that

$$\text{Ad}(\tilde{A})^* (\phi_j^i(\zeta)) = [\phi_j^i \circ \text{Ad}(A)](\zeta).$$

Using equation (1), we find that for all $f: \mathfrak{o}(n) \rightarrow \mathbb{R}$ we have

$$(*) \quad \text{Ad}(\tilde{A})^* f(\zeta) = [f \circ \text{Ad}(A)](\zeta).$$

From equation (*) we see that a linear combination $f(\zeta)$ is invariant under all $\text{Ad}(\tilde{A})^*$, and thus $f(\zeta) \in \Omega^k(\mathfrak{h}^\perp)^H \approx H^k(\tilde{G}_n(\mathbb{R}^N))$, if and only if $f: \mathfrak{o}(n) \rightarrow \mathbb{R}$ is invariant under all $\text{Ad}(A)$, for $A \in \text{SO}(n)$. But Theorem 40 says that all such f are polynomials in

$$\begin{aligned} f_2, \dots, f_{2[n/2]} & \quad \text{if } n \text{ is odd} \\ f_2, \dots, f_{n-2}, \text{Pf} & \quad \text{if } n \text{ is even.} \end{aligned}$$

Moreover, f is uniquely expressible as such a polynomial, since $f_2, \dots, f_{2[n/2]}$ [or $f_2, \dots, f_{n-2}, \text{Pf}$] are algebraically independent (pages 334 and 336).

Case 1. n is odd. The forms $f_2(\zeta), \dots, f_{2[n/2]}(\zeta)$ have dimensions $4, 8, \dots, 4[n/2]$. So

If $k < N - n$ is not a multiple of 4, then $H^k(\tilde{G}_n(\mathbb{R}^N)) = 0$.

If $k < N - n$ is a multiple of 4, then every element of $H^k(\tilde{G}_n(\mathbb{R}^N))$ is a unique linear combination of cup products of the classes corresponding, via Theorem 32, to the forms

$$f_2(\zeta), \dots, f_{k/4}(\zeta).$$

Case 2. n is even. The forms $f_2(\zeta), \dots, f_{n-2}(\zeta), \text{Pf}(\zeta)$ have dimensions $4, 8, \dots, 2n - 4, n$. So

If $k < N - n$ is odd, then $H^k(\tilde{G}_n(\mathbb{R}^N)) = 0$.

If $k < N - n$ is even, then every element of $H^k(\tilde{G}_n(\mathbb{R}^N))$ is a unique linear combination of cup products of the classes corresponding, via Theorem 32, to the forms

$$f_2(\zeta), \dots, f_{[k/4]}(\zeta), \quad \text{and} \quad \text{Pf}(\zeta) \quad \text{if } k \geq n.$$

This can all be said more prettily if we fix n and allow N to increase:

41. PROPOSITION. If $\alpha: \tilde{G}_n(\mathbb{R}^N) \rightarrow \tilde{G}_n(\mathbb{R}^M)$ is the natural map, and we have $M > N > n + k$, then the induced map

$$\alpha^*: H^k(\tilde{G}_n(\mathbb{R}^M)) \rightarrow H^k(\tilde{G}_n(\mathbb{R}^N))$$

is an isomorphism.

PROOF. Because of the preceding discussion, it obviously suffices to show that the element of $H^k(\tilde{G}_n(\mathbb{R}^M))$ corresponding to f_r goes by α^* to the element in $H^k(\tilde{G}_n(\mathbb{R}^N))$ corresponding to f_r . Proving this is just a matter of unraveling definitions, and will provide a good opportunity to set straight everything done up till now. ♦

Henceforth we consider only N sufficiently large so that $4[n/2] < N - n$ for odd n , and $2n - 4 < N - n$ and $n < N - n$ for even n (we can take $N > 3n - 2$ in both cases). Then all elements of $H^*(\tilde{G}_n(\mathbb{R}^N))$ in dimensions $< N - n$ are unique linear combinations of cup products of the classes corresponding to

$$\begin{array}{ll} f_2(\zeta), \dots, f_{2[n/2]}(\zeta) & n \text{ odd} \\ f_2(\zeta), \dots, f_{n-2}(\zeta), \text{Pf}(\zeta) & n \text{ even.} \end{array}$$

We let

$$p_{n;k} \in H^{4k}(\tilde{G}_n(\mathbb{R}^N)) \quad k = 1, \dots, [n/2]$$

be the class corresponding to

$$\frac{1}{(2\pi)^{2k}} f_{2k}(\zeta),$$

and we let

$$e_n \in H^n(\tilde{G}_n(\mathbb{R}^N)) \quad n = 2m$$

be the class corresponding to

$$\frac{1}{(2\pi)^m} \text{Pf}(\zeta).$$

We defined $p_{n;k}$ for $k = 1, \dots, [n/2]$ for both odd and even n , just for simplicity. For even n this gives us the extra class $p_{n;n/2}$, corresponding to $f_n(\zeta)/(2\pi)^n$. It satisfies

$$p_{n;n/2} = e_n \cup e_n,$$

since $e_n \cup e_n$ corresponds to

$$\begin{aligned} \frac{1}{(2\pi)^m} \text{Pf}(\zeta) \wedge \frac{1}{(2\pi)^m} \text{Pf}(\zeta) &= \frac{1}{(2\pi)^n} \text{Pf}^2(\zeta) \quad \text{by equation (1) on page 341} \\ &= \frac{1}{(2\pi)^n} \det(\zeta) = \frac{1}{(2\pi)^n} f_n(\zeta). \end{aligned}$$

In these definitions, we are always taking N large, and applying Proposition 41, so that there is no need to have an extra subscript N on the symbols $p_{n;k}$ and e_n .

In accordance with our discussion in section 5, each class

$$p_{n;k} \in H^{4k}(\tilde{G}_n(\mathbb{R}^N)) \quad N \text{ large}$$

determines a “characteristic class”, that is, a function

$$\xi \mapsto p_{n;k}(\xi) \in H^{4k}(M)$$

which assigns to a smooth oriented n -dimensional bundle $\xi = \pi: E \rightarrow M$ an element of the cohomology of M . Explicitly,

$$p_{n;k}(\xi) = g^* p_{n;k}$$

where

$$g: M \rightarrow \tilde{G}_n(\mathbb{R}^N) \quad \text{satisfies} \quad g^* \tilde{\gamma}^n(\mathbb{R}^N) \simeq \xi.$$

This characteristic class is called the k^{th} **Pontryagin class** for n -dimensional bundles. For even n , we have the additional class

$$\xi \mapsto e_n(\xi).$$

When one is dealing with characteristic classes, the number n is usually apparent, since it is the fibre dimension of the bundle whose characteristic class is being considered. Consequently, we write simply $p_k(\xi)$ and $e(\xi)$. If ξ has fibre dimension n , then $p_k(\xi)$ is defined for $k = 1, \dots, [n/2]$; if n is even, then we also have the class $e(\xi)$, and $p_{n/2}(\xi) = e(\xi) \cup e(\xi)$.

Since all elements of $H^*(\tilde{G}_n(\mathbb{R}^N))$ in dimensions $< N - n$ are linear combinations of the $p_{n;k}$ and e_n , we see that *all* characteristic classes for oriented n -dimensional bundles are polynomials in the Pontryagin classes p_k , together with e if n is even. In particular, the Euler class must be representable in this way, and our notation clearly suggests that the Euler class is, in fact, just the characteristic class e . In order to prove this, we have to look a little more carefully at the universal bundles.

Consider the universal bundle $\tilde{\gamma}^n(\mathbb{R}^N) = \pi: E(\tilde{\gamma}^n(\mathbb{R}^N)) \rightarrow \tilde{G}_n(\mathbb{R}^N)$. A point of $\tilde{G}_n(\mathbb{R}^N)$ is an oriented n -dimensional subspace $W \subset \mathbb{R}^N$, and the fibre $\pi^{-1}(W)$ over W is $\{(W, w) : w \in W\}$. So there is a natural Riemannian metric $\langle \cdot, \cdot \rangle$ on $\tilde{\gamma}^n(\mathbb{R}^N)$: the inner product of $(W, w_1), (W, w_2) \in \pi^{-1}(W)$ is just the usual inner product of $w_1, w_2 \in \mathbb{R}^N$. For the corresponding principal bundle $\varpi: \text{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N))) \rightarrow \tilde{G}_n(\mathbb{R}^N)$, the fibre $\varpi^{-1}(W)$ is the set of all $(W, (w_1, \dots, w_n))$, where (w_1, \dots, w_n) is a positively oriented orthonormal n -frame in $W \subset \mathbb{R}^N$. Now we can define a map $\lambda: \text{SO}(N) \rightarrow \text{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N)))$,

from the special orthogonal group $\mathrm{SO}(N)$ to the total space of this principal bundle, as follows: If $W_0 \subset \mathbb{R}^N$ is the subspace spanned by e_1, \dots, e_n , then

$$\lambda(A) = (A(W_0), (A(e_1), \dots, A(e_n))).$$

It is easy to see that for the point $x = (W_0, (e_1, \dots, e_n)) \in \mathrm{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N)))$, we have

$$\lambda^{-1}(x) = \{I\} \times \mathrm{SO}(N - n);$$

more generally, for any $x \in \mathrm{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N)))$ the set $\lambda^{-1}(x)$ is a left coset of $\{I\} \times \mathrm{SO}(N - n)$. So $\mathrm{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N)))$ can be identified with the left coset space

$$\mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N - n).$$

We leave it as an exercise for the reader to show that the topology and C^∞ structure on $\mathrm{SO}(E(\tilde{\gamma}^n(\mathbb{R}^N)))$ is the same as that on this left coset space, and that the projection

$$\varpi: \mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N - n) \rightarrow \mathrm{SO}(N)/\mathrm{SO}(n) \times \mathrm{SO}(N - n)$$

is just the natural map taking the coset $A \cdot [\{I\} \times \mathrm{SO}(N - k)]$ to the coset $A \cdot [\mathrm{SO}(k) \times \mathrm{SO}(N - k)]$. Notice that the diagram

$$\begin{array}{ccc} \mathrm{SO}(N) & \xrightarrow{\pi_1} & \frac{\mathrm{SO}(N)}{\{I\} \times \mathrm{SO}(N - n)} \\ & \searrow \pi_2 & \downarrow \varpi \\ & & \frac{\mathrm{SO}(N)}{\mathrm{SO}(n) \times \mathrm{SO}(N - n)} \end{array}$$

commutes, where π_1 and π_2 are the natural projections. As in section 6, we will use L_A for the left multiplication $L_A: \mathrm{SO}(N) \rightarrow \mathrm{SO}(N)$, and \mathbf{L}_A for the diffeomorphism of $\mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N - n)$ taking the coset $B \cdot [\{I\} \times \mathrm{SO}(N - n)]$ to $AB \cdot [\{I\} \times \mathrm{SO}(N - n)]$. We also have the map $R_A: \mathrm{SO}(N) \rightarrow \mathrm{SO}(N)$, and the map \mathbf{R}_A taking the coset $B \cdot [\{I\} \times \mathrm{SO}(N - n)]$ to $BA \cdot [\{I\} \times \mathrm{SO}(N - n)]$. The reader should check that for $A \in \mathrm{SO}(n)$, the map \mathbf{R}_A corresponds to the right multiplication by A in the principal bundle $\mathrm{SO}(\tilde{\gamma}^n(\mathbb{R}^N))$.

On $\mathrm{SO}(N)$ we have the left invariant 1-forms $\tilde{\phi}_\alpha^\beta$ ($\alpha < \beta$) whose values at $I \in \mathrm{SO}(N)$ are the elements $\phi_\alpha^\beta \in \mathfrak{o}(N)^*$; set $\tilde{\phi}_\alpha^\beta = -\tilde{\phi}_\beta^\alpha$ for $\alpha > \beta$, and $\tilde{\phi}_\alpha^\alpha = 0$. We claim that for $i, j \leq n$ there are unique 1-forms ω_j^i on $\mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N - n)$ such that

$$(1) \quad \pi_1^* \omega_j^i = \tilde{\phi}_j^i.$$

To prove this we first note that π_{1*} is always onto. Since we need to have

$$(2) \quad \omega_j^i(\pi_{1*}X) = \tilde{\phi}_j^i(X)$$

for all tangent vectors X of $\mathrm{SO}(N)$, this proves uniqueness. To prove existence, we need to show that definition (2) is well-defined, by showing that $\tilde{\phi}_j^i(X) = 0$ whenever $\pi_{1*}X = 0$. So suppose $X \in \mathrm{SO}(N)_A$. Then $X = L_{A*}X_I$ for some $X_I \in \mathfrak{o}(N)$. Since $\pi_1 \circ L_A = \mathbf{L}_A \circ \pi_1$, we see that

$$\begin{aligned} \pi_{1*}X = 0 &\implies \pi_{1*}L_{A*}X_I = 0 \\ &\implies \mathbf{L}_{A*}\pi_{1*}X_I = 0 \\ &\implies \pi_{1*}X_I = 0 \quad \text{since } \mathbf{L}_A \text{ is a diffeomorphism} \\ &\implies X_I \text{ is of the form } \begin{pmatrix} 0 & 0 \\ 0 & * \end{pmatrix} \\ &\implies \phi_j^i(X_I) = 0 \\ &\implies \tilde{\phi}_j^i(X) = 0 \quad \text{since } \tilde{\phi}_j^i \text{ is left invariant.} \end{aligned}$$

Thus the forms ω_j^i exist. Note that

$$\begin{aligned} \mathbf{L}_A^*\omega_j^i(\pi_{1*}X) &= \omega_j^i(\mathbf{L}_{A*}\pi_{1*}X) \\ &= \omega_j^i(\pi_{1*}L_{A*}X) \\ &= \tilde{\phi}_j^i(L_{A*}X) \\ &= \tilde{\phi}_j^i(X) = \omega_j^i(\pi_{1*}X). \end{aligned}$$

So

$$\mathbf{L}_A^*\omega_j^i = \omega_j^i.$$

Now $\omega = (\omega_j^i)$ is an $\mathfrak{o}(n)$ -valued 1-form on $\mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N-n)$. We claim that ω is, in fact, a connection on the principal bundle $\varpi: \mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N-n) \rightarrow \mathrm{SO}(N)/\mathrm{SO}(n) \times \mathrm{SO}(N-n)$. We have to check that

$$\begin{aligned} \omega(\sigma(M)) &= M && \text{for } M \in \mathfrak{o}(n) \\ \omega(\mathbf{R}_{A*}Y) &= A^{-1}\omega(Y)A && \text{for } A \in \mathrm{SO}(n) \text{ and } Y \text{ a tangent} \\ &&& \text{vector on } \mathrm{SO}(N)/\{I\} \times \mathrm{SO}(N-n). \end{aligned}$$

Recall that the value of $\sigma(M)$ at the coset $B \cdot [\{I\} \times \mathrm{SO}(N-n)]$ is $c'(0)$ where

$$\begin{aligned} c(t) &= \mathbf{R}_{\exp t M}(B \cdot [\{I\} \times \mathrm{SO}(N-n)]) \\ &= B(\exp t M) \cdot [\{I\} \times \mathrm{SO}(N-n)] \\ &= \pi_1 L_B(\exp t M) \\ \implies c'(0) &= \pi_{1*}L_{B*}M. \end{aligned}$$

Thus

$$\begin{aligned}
 \omega_j^i(\sigma(M) \text{ at } B \cdot [\{I\} \times \text{SO}(N-n)]) &= \omega_j^i(c'(0)) \\
 &= \omega_j^i(\pi_{1*}L_{B*}M) \\
 &= \tilde{\omega}_j^i(L_{B*}M) \\
 &= \phi_j^i(M) = M_j^i,
 \end{aligned}$$

which proves the first condition. To prove the second, take a tangent vector Y at the coset $B \cdot [\{I\} \times \text{SO}(N-n)]$ and choose a tangent vector $X \in \mathfrak{o}(N)_B$ with $\pi_{1*}X = Y$. Then

$$\omega_j^i(Y) = \omega_j^i(\pi_{1*}X) = \tilde{\phi}_j^i(X) = \phi_j^i(L_{B^{-1}*}X),$$

while

$$\begin{aligned}
 \omega_j^i(\mathbf{R}_{A*}Y) &= \omega_j^i(\mathbf{R}_{A*}\pi_{1*}X) = \omega_j^i(\pi_{1*}R_{A*}X) \\
 &= \tilde{\phi}_j^i(R_{A*}X) && R_{A*}X \text{ a tangent vector at } BA \\
 &= \phi_j^i(L_{(BA)^{-1}*}R_{A*}X) \\
 &= \phi_j^i(L_{A^{-1}*}R_{A*}L_{B^{-1}*}X) \\
 &= \phi_j^i(\text{Ad}(A^{-1})L_{B^{-1}*}X) \\
 &= \phi_j^i(A^{-1}(L_{B^{-1}*}X)A) \\
 &= \sum_{\mu, \nu=1}^k (A^{-1})_{\mu}^i \phi_j^i(L_{B^{-1}*}X)_{\nu}^{\mu} A_j^{\nu} \quad \text{by linearity of } \phi_j^i,
 \end{aligned}$$

which proves the second condition.

It is easy to see which vectors $\pi_{1*}X$ are vertical or horizontal when $X \in \mathfrak{o}(N)$. First of all,

$$\begin{aligned}
 \pi_{1*}X \text{ is vertical} &\iff \varpi_*\pi_{1*}X = 0 \\
 &\iff \pi_{2*}X = 0 \\
 &\iff X \in \mathfrak{o}(n) \times \mathfrak{o}(N-n).
 \end{aligned}$$

On the other hand,

$$\begin{aligned}
 \pi_{1*}X \text{ is horizontal} &\iff \text{all } \omega_j^i(\pi_{1*}X) = 0 \\
 &\iff \text{all } \omega_j^i(X) = 0 \\
 &\iff X \text{ has the form } \begin{pmatrix} 0 & 0 \\ 0 & * \end{pmatrix}.
 \end{aligned}$$

Given $X \in \mathfrak{o}(N)$, we write it as

$$\begin{aligned} X &= \begin{pmatrix} 0 & * \\ * & * \end{pmatrix} + \begin{pmatrix} * & 0 \\ 0 & 0 \end{pmatrix} \\ &= X_1 + X_2. \end{aligned}$$

Then $\pi_{1*}X_1$ is horizontal and $\pi_{1*}X_2$ is vertical. This means that the horizontal component of $\pi_{1*}X$ is precisely

$$h(\pi_{1*}X) = \pi_{1*}X_1.$$

To compute the curvature forms Ω_j^i for the connection ω_j^i , we use the fact (Problem 7-15) that the forms $\tilde{\phi}_\alpha^\beta$ satisfy

$$d\tilde{\phi}_\alpha^\beta = - \sum_{\gamma=1}^N \tilde{\phi}_\gamma^\beta \wedge \tilde{\phi}_\alpha^\gamma.$$

Then for $X, Y \in \mathfrak{o}(N)$ we have

$$\begin{aligned} \Omega_j^i(\pi_{1*}X, \pi_{1*}Y) &= d\omega_j^i(h\pi_{1*}X, h\pi_{1*}Y) \\ &= d\omega_j^i(\pi_{1*}X_1, \pi_{1*}Y_1) \\ &= d(\pi_1^*\omega_j^i)(X_1, Y_1) \\ &= d\tilde{\phi}_j^i(X_1, Y_1) \\ &= - \sum_{\gamma=1}^N \phi_\gamma^i \wedge \phi_j^\gamma(X_1, Y_1). \end{aligned}$$

Since

$$X_1 = X - \sum_{\alpha \text{ or } \beta > n} \phi_\alpha^\beta(X) \cdot X_\alpha^\beta, \quad Y_1 = Y - \sum_{\alpha \text{ or } \beta > n} \phi_\alpha^\beta(Y) \cdot X_\alpha^\beta,$$

this gives

$$\begin{aligned} \Omega_j^i(\pi_{1*}X, \pi_{1*}Y) &= \sum_r \phi_i^r \wedge \phi_j^r(X, Y) \\ &= \zeta_{ij}(X, Y), \end{aligned}$$

which shows that

$$(*) \quad \pi_1^*\Omega_j^i = \zeta_{ij} \quad \text{at } \text{SO}(N)_I.$$

(In fact, we also have $\pi_1^*\Omega_j^i = \tilde{\zeta}_{ij}$, where $\tilde{\zeta}_{ij}$ is the left invariant form extending ζ_{ij} , since the equation $\mathbf{L}_A^*\omega_j^i = \omega_j^i$ implies that we also have $\mathbf{L}_A^*\Omega_j^i = \Omega_j^i$.)

42. THEOREM. If ξ is an oriented n -dimensional bundle, with $n = 2m$ even, then the characteristic class $e(\xi)$ is the Euler class $\chi(\xi)$.

PROOF. It suffices to prove this when ξ is the universal bundle $\tilde{\gamma}^n(\mathbb{R}^N)$. We know, by Corollary 25 and Theorem 26, that the Euler class $\chi(\tilde{\gamma}^n(\mathbb{R}^N))$ is represented by the unique form Γ on $\text{SO}(N)/\text{SO}(n) \times \text{SO}(N - n)$ such that

$$\begin{aligned}\varpi^* \Gamma &= \frac{1}{\pi^m m! 2^n} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}} \\ &= \frac{1}{\pi^m m! 2^n} 2^m \cdot m! \text{Pf}(\Omega) \\ &= \frac{1}{(2\pi)^m} \text{Pf}(\Omega).\end{aligned}$$

We want to show that Γ corresponds, via Theorem 32, to the form

$$\frac{1}{(2\pi)^m} \text{Pf}(\zeta).$$

Note first that if L'_A is the diffeomorphism of $\text{SO}(N)/\text{SO}(n) \times \text{SO}(N - n)$ taking the coset $B \cdot [\text{SO}(n) \times \text{SO}(N - n)]$ to $AB \cdot [\text{SO}(n) \times \text{SO}(N - n)]$, then $L'_A \circ \varpi = \varpi \circ L_A$. So

$$\begin{aligned}\varpi^*(L'_A{}^* \Gamma) &= L_A{}^* \varpi^* \Gamma = \frac{1}{(2\pi)^m} L_A{}^* \text{Pf}(\Omega) \\ &= \frac{1}{(2\pi)^m} \text{Pf}(L_A{}^* \Omega) = \frac{1}{(2\pi)^m} \text{Pf}(\Omega) \\ &= \varpi^* \Gamma.\end{aligned}$$

By uniqueness in Proposition 18 we have $L'_A{}^* \Gamma = \Gamma$ for all $A \in \text{SO}(N)$. In other words, Γ is an invariant form on $\text{SO}(N)/\text{SO}(n) \times \text{SO}(N - n)$, as defined in section 6. So the element of $\Omega^n(\mathfrak{o}(N))$ corresponding to Γ in Theorem 32 is simply $\pi_2^* \Gamma$ at I . But

$$\begin{aligned}\pi_2^* \Gamma &= \pi_1^* \varpi^* \Gamma \\ &= \frac{1}{(2\pi)^m} \pi_1^* \text{Pf}(\Omega) = \frac{1}{(2\pi)^m} \text{Pf}(\pi_1^* \Omega) \\ &= \frac{1}{(2\pi)^m} \text{Pf}(\zeta) \quad \text{at } I, \quad \text{by equation (*).} \quad \blacklozenge\end{aligned}$$

The classes $p_k(\xi)$ of an oriented bundle $\xi = \pi: E \rightarrow M$ may be described in exactly the same way: Choose a Riemannian metric $\langle \cdot, \cdot \rangle$ for ξ , form the principal bundle $O(\xi) = \varpi: O(E) \rightarrow M$, and let ω be a connection on $O(E)$, with curvature form Ω .

43. THEOREM. There is a unique $4k$ -form Λ on M such that

$$\varpi^* \Lambda = \frac{1}{(2\pi)^{2k}} f_{2k}(\Omega).$$

This form Λ is closed, and the cohomology class $[\Lambda]$ is independent of the choice of the Riemannian metric $\langle \cdot, \cdot \rangle$ and the connection ω in terms of which Λ is defined. The cohomology class $[\Lambda]$ is precisely $p_k(\xi)$.

PROOF. The proofs of Propositions 18, 19, and 20 can be adapted, essentially without modification, to prove the first two assertions. The proof of the final assertion is exactly like the proof of Theorem 42. ♦

As a simple application, we consider a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ of constant curvature K_0 . Then the curvature form Ω on $O(TM)$ satisfies

$$\Omega_j^i = K_0 \theta^i \wedge \theta^j.$$

To calculate $f_{2k}(\Omega)$, we use the explicit formula given in Problem I.7-14, to obtain

$$\begin{aligned} f_{2k}(\Omega) &= \frac{1}{(2k)!} \sum_{\substack{i_1 \dots i_{2k} \\ j_1 \dots j_{2k}}} \Omega_{i_1}^{j_1} \wedge \dots \wedge \Omega_{i_{2k}}^{j_{2k}} \delta_{j_1 \dots j_{2k}}^{i_1 \dots i_{2k}} \\ &= \frac{(K_0)^{2k}}{(2k)!} \sum_{\substack{i_1 \dots i_{2k} \\ j_1 \dots j_{2k}}} \theta^{j_1} \wedge \theta^{i_1} \wedge \dots \wedge \theta^{j_{2k}} \wedge \theta^{i_{2k}} \delta_{j_1 \dots j_{2k}}^{i_1 \dots i_{2k}}. \end{aligned}$$

In this sum, the δ term vanishes unless j_1, \dots, j_{2k} is a permutation of i_1, \dots, i_{2k} ; but then $\theta^{j_1} \wedge \dots \wedge \theta^{i_{2k}}$ has repeated factors, so it vanishes. Thus,

44. COROLLARY. If M^n is a compact manifold of constant curvature, then

$$p_k(TM) = 0, \quad k = 1, \dots, [n/4].$$

Another application of Theorem 43 gives us an analogue of Theorem 17. For a bundle ξ over M we define the **total Pontryagin class** $p(\xi)$ to be the element of $H^0(M) \oplus H^4(M) \oplus \dots$ given by

$$p(\xi) = 1 + p_1(\xi) + \dots + p_{[n/2]}(\xi) = p_0(\xi) + p_1(\xi) + \dots + p_{[n/2]}(\xi),$$

where $1 \in H^0(M)$ is the standard element (represented by the constant function 1 on M).

45. THEOREM. If ξ and η are oriented bundles over the same compact manifold M , then the total Pontryagin class of $\xi \oplus \eta$ is given by the **Whitney product formula**

$$p(\xi \oplus \eta) = p(\xi) \cup p(\eta).$$

[This means that

$$p_k(\xi \oplus \eta) = \sum_{l=0}^k p_l(\xi) \cup p_{k-l}(\eta),$$

when we look at individual components.]

PROOF. The proof will be almost exactly like the proof of Theorem 22. For convenience we rename our bundles ξ and η as $\xi_i = \pi_i: E_i \rightarrow M$ for $i = 1, 2$, and let $\xi_1 \oplus \xi_2 = \pi: E \rightarrow M$. We introduce the corresponding principal bundles $\varpi_i: \text{SO}(E_i) \rightarrow M$ and $\varpi: \text{SO}(E) \rightarrow M$, the principal bundle $\text{SO}(E_1) * \text{SO}(E_2) \rightarrow M$, and the projections $\rho_i: \text{SO}(E_1) * \text{SO}(E_2) \rightarrow \text{SO}(E_i)$. Choose connections ω_i on $\text{SO}(E_i)$, with curvature forms Ω_i . Then

$$\rho_1^* \omega_1 \oplus \rho_2^* \omega_2 = \begin{pmatrix} \rho_1^* \omega_1 & 0 \\ 0 & \rho_2^* \omega_2 \end{pmatrix}$$

is a connection $\bar{\omega}$ on $\text{SO}(E_1) * \text{SO}(E_2)$, with curvature form

$$\bar{\Omega} = \rho_1^* \Omega_1 \oplus \rho_2^* \Omega_2 = \begin{pmatrix} \rho_1^* \Omega_1 & 0 \\ 0 & \rho_2^* \Omega_2 \end{pmatrix},$$

and $\bar{\omega}$ can be extended uniquely to a connection $\tilde{\omega}$ on $\text{SO}(E)$. At a point $e \in \text{SO}(E_1) * \text{SO}(E_2)$ we have

$$\tilde{\Omega} = \bar{\Omega} \quad (\text{on tangent vectors to } \text{SO}(E_1) * \text{SO}(E_2))$$

which implies that

$$\begin{aligned} f_{2k}(\tilde{\Omega}) &= f_{2k}(\bar{\Omega}) = \sum_{l=0}^k f_{2l}(\rho_1^* \Omega_1) \wedge f_{2k-2l}(\rho_2^* \Omega_2) \\ &\quad \text{by the formula on page 337} \\ &= \sum_{l=0}^k \rho_1^* f_{2l}(\Omega_1) \wedge \rho_2^* f_{2k-2l}(\Omega_2). \end{aligned}$$

So if Λ_k is the form representing $p_{2k}(\xi_1 \oplus \xi_2)$ while Υ_l^i are the forms representing $p_{2l}(\xi_i)$, then at e we have [on tangent vectors to $\mathrm{SO}(E_1) * \mathrm{SO}(E_2)$]

$$\begin{aligned} \varpi^* \Lambda_k &= \frac{1}{(2\pi)^{2k}} f_{2k}(\tilde{\Omega}) = \sum_{l=0}^k \frac{1}{(2\pi)^l} \rho_1^* f_{2l}(\Omega_1) \wedge \frac{1}{(2\pi)^{2k-l}} \rho_2^* f_{2k-2l}(\Omega_2) \\ &= \sum_{l=0}^k \rho_1^* \varpi_1^* \Upsilon_l^1 \wedge \rho_2^* \varpi_2^* \Upsilon_{k-l}^2 \\ &= \sum_{l=0}^k \varpi^* \Upsilon_l^1 \wedge \varpi^* \Upsilon_{k-l}^2. \end{aligned}$$

This implies that

$$\Lambda_k = \sum_{l=0}^k \Upsilon_l^1 \wedge \Upsilon_{k-l}^2. \quad \spadesuit$$

10. THE WEIL HOMOMORPHISM

The invariant polynomial functions $f_{2k}: \mathfrak{o}(n) \rightarrow \mathbb{R}$ and $\mathrm{Pf}: \mathfrak{o}(n) \rightarrow \mathbb{R}$ arose naturally in our attempts to calculate the cohomology of $\tilde{G}_n(\mathbb{R}^N)$; each one gave us an element of $H^*(\tilde{G}_n(\mathbb{R}^N))$, and hence a characteristic class $\xi \mapsto C(\xi)$ for oriented bundles. On the other hand, at the end of the last section we saw how these characteristic classes $C(\xi)$ could be defined directly for the bundle ξ , by means of a connection on the associated principal bundle $\mathrm{SO}(\xi)$. There is no reason why we cannot use exactly the same procedure for groups other than $\mathrm{SO}(n)$.

For any Lie group G , with Lie algebra \mathfrak{g} , we consider the set $\mathcal{P}(\mathfrak{g})$ of functions $f: \mathfrak{g} \rightarrow \mathbb{R}$ which can be expressed as polynomials in $\{\phi^\alpha\}$, where $\{\phi^\alpha\}$ is a basis of \mathfrak{g}^* . Such functions are called **polynomial functions** on \mathfrak{g} (the concept is clearly independent of the choice of basis $\{\phi_\alpha\}$), and the set of all homogeneous polynomial functions of degree k will be denoted by $\mathcal{P}^k(\mathfrak{g})$. We say that $f: \mathfrak{g} \rightarrow \mathbb{R}$ is **Ad(G)-invariant** if $f \circ \mathrm{Ad}(a) = f$ for all $a \in G$. Instead of considering polynomial functions on \mathfrak{g} it is often more convenient to consider the set $\mathcal{S}^k(\mathfrak{g})$ of symmetric k -linear maps $f: \mathfrak{g} \times \cdots \times \mathfrak{g} \rightarrow \mathbb{R}$. Given $f \in \mathcal{S}^k(\mathfrak{g})$, we define a polynomial function $\mathcal{P}f \in \mathcal{P}^k(\mathfrak{g})$ by

$$(\mathcal{P}f)(X) = f(X, \dots, X) \quad X \in \mathfrak{g}.$$

Conversely, given a basis ϕ^1, \dots, ϕ^r of \mathfrak{g} , and a polynomial function f of degree k on \mathfrak{g} , we can write it uniquely as

$$\sum_{\alpha_1, \dots, \alpha_k=1}^r a_{\alpha_1 \dots \alpha_k} \phi^{\alpha_1} \dots \phi^{\alpha_k}$$

where the $a_{\alpha_1 \dots \alpha_k}$ are symmetric in $\alpha_1, \dots, \alpha_k$; then we define $\mathcal{J}f \in \mathcal{J}^k(\mathfrak{g})$ by

$$(\mathcal{J}f)(X_1, \dots, X_k) = \sum a_{\alpha_1 \dots \alpha_k} \phi^{\alpha_1}(X_1) \dots \phi^{\alpha_k}(X_k), \quad X_1, \dots, X_k \in \mathfrak{g}.$$

It is easy to check that the maps

$$\mathcal{P}: \mathcal{J}^k(\mathfrak{g}) \rightarrow \mathcal{P}^k(\mathfrak{g}), \quad \mathcal{J}: \mathcal{P}^k(\mathfrak{g}) \rightarrow \mathcal{J}^k(\mathfrak{g})$$

are inverses to each other (so \mathcal{J} doesn't depend on the choice of basis). For $f \in \mathcal{J}^k(\mathfrak{g})$ and $g \in \mathcal{J}^l(\mathfrak{g})$ we define $fg \in \mathcal{J}^{k+l}(\mathfrak{g})$ by

$$\begin{aligned} fg(X_1, \dots, X_{k+l}) \\ = \frac{1}{(k+l)!} \sum_{\pi \in S_{k+l}} f(X_{\pi(1)}, \dots, X_{\pi(k)}) \cdot g(X_{\pi(k+1)}, \dots, X_{\pi(k+l)}). \end{aligned}$$

This makes $\mathcal{P}(fg) = \mathcal{P}(f) \cdot \mathcal{P}(g)$. We define $f \in \mathcal{J}^k(\mathfrak{g})$ to be **Ad(G)-invariant** if

$$\begin{aligned} f(\text{Ad}(a)X_1, \dots, \text{Ad}(a)X_k) &= f(X_1, \dots, X_k) \\ \text{for all } a \in G \text{ and } X_1, \dots, X_k \in \mathfrak{g}; \end{aligned}$$

then f is Ad(G)-invariant if and only if $\mathcal{P}f$ is Ad(G)-invariant. The set of all $f \in \mathcal{J}^k(\mathfrak{g})$ which are Ad(G)-invariant is denoted by $I^k(G)$. Thus, \mathcal{P} takes $I^k(G)$ into the set of polynomial functions on \mathfrak{g} which are Ad(G)-invariant, and \mathcal{J} takes this set back to $I^k(G)$.

Now let $\pi: P \rightarrow M$ be a principal bundle with group G , and let ω be a connection, with curvature form Ω . Thus both ω and Ω are \mathfrak{g} -valued, so if ϕ^1, \dots, ϕ^r is a basis of \mathfrak{g}^* , then we can write $\omega = \sum \omega^\alpha \cdot \phi^\alpha$ and $\Omega = \sum \Omega^\alpha \cdot \phi^\alpha$ for ordinary forms ω^α and Ω^α . Given $f \in \mathcal{P}^k(\mathfrak{g})$, we write it as a sum of terms

$$c \cdot \phi^{\alpha_1} \dots \phi^{\alpha_k},$$

and then let $f(\Omega)$ be the $2k$ -form on P which is the corresponding sum of terms

$$c \cdot \Omega^{\alpha_1} \wedge \dots \wedge \Omega^{\alpha_k}$$

(since the Ω^α are 2-forms, the order in the product $\phi^{\alpha_1} \dots \phi^{\alpha_k}$ is irrelevant). This is the definition used previously for the case $G = \mathrm{SO}(n)$, where the Lie algebra $\mathfrak{o}(n)$ has a natural basis $\{\phi_j^i\}_{i < j}$ (provided we use the convention that a polynomial in the ϕ_j^i ($i, j = 1, \dots, n$) be interpreted as a polynomial in the $\{\phi_j^i\}_{i < j}$ by replacing ϕ_i^j by $-\phi_j^i$ for $j > i$). A more intrinsic description is possible when we work with $f \in \mathcal{J}^k(\mathfrak{g})$. We now define $f(\Omega)$ to be the $2k$ -form on P given by

$$\begin{aligned} f(\Omega)(X_1, \dots, X_{2k}) \\ = \frac{1}{2^k} \sum_{\pi \in S_{2k}} (\mathrm{sgn} \pi) \cdot f(\Omega(X_{\pi(1)}, X_{\pi(2)}), \dots, \Omega(X_{\pi(2k-1)}, X_{\pi(2k)})), \end{aligned}$$

where X_1, \dots, X_{2k} are now tangent vectors of P . It can be checked that $(\mathcal{P}f)(\Omega) = f(\Omega)$ for $f \in \mathcal{J}^k(\mathfrak{g})$; equivalently, $(\mathcal{J}f)(\Omega) = f(\Omega)$ for $f \in \mathcal{P}^k(\mathfrak{g})$ (so the definition of $f(\Omega)$ doesn't depend on the choice of $\{\phi^\alpha\}$).

46. THEOREM. Let $\xi = \pi: P \rightarrow M$ be a principal bundle with group G , and let ω be a connection on P , with curvature form Ω . Then for every $f \in I^k(G)$ there is a unique $2k$ -form Λ on M such that

$$\pi^* \Lambda = f(\Omega).$$

The form Λ is closed, and its de Rham cohomology class $w_\xi(f) = [\Lambda]$ is independent of the choice of ω . For $f \in I^k(G)$ and $g \in I^l(G)$ we have $w_\xi(fg) = w_\xi(f) \cup w_\xi(g)$.

PROOF. Exactly like the proofs of Theorems 18, 19, and 20. ♦

If we set $H^*(M) = H^0(M) \oplus H^1(M) \oplus \dots$ and $I(G) = \mathbb{R} \oplus I^1(G) \oplus \dots$, then we have a homomorphism $w_\xi: I(G) \rightarrow H^*(M)$, depending only on the given principal bundle $\xi = \pi: P \rightarrow M$. This map is called the **Weil homomorphism**. It is natural, in the following sense.

47. PROPOSITION. Let $\pi: P \rightarrow M$ be a principal bundle with group G and let $f: M' \rightarrow M$ be a smooth map, inducing the map $f^*: H^*(M) \rightarrow H^*(M')$. Then

$$w_{f^*\xi} = f^* \circ w_\xi.$$

PROOF. An elementary exercise (just like the proof of Proposition 21). ♦

If we take $G = \mathrm{SO}(n)$ in Theorem 46, and consider the functions $g_{2k} = \delta(f_{2k}) \in I^{2k}(\mathrm{SO}(n))$ corresponding to the polynomial functions f_{2k} on $\mathfrak{o}(n)$, then we have classes $w_\eta(g_{2k}) \in H^{4k}(M)$ for any principal $\mathrm{SO}(n)$ bundle η over M . If $\xi = \pi: E \rightarrow M$ is an oriented n -dimensional vector bundle over M , then we can form the principal bundle $\eta = \mathrm{SO}(\xi)$ by means of a Riemannian metric on ξ [all such η are equivalent by Corollary 5], and Theorem 43 amounts to the assertion that $w_\eta(g_{2k}) = p_k(\xi)$. Notice that the $\mathrm{SO}(n)$ -invariant polynomials f_{2k} on $\mathfrak{o}(n)$ are also $\mathrm{GL}(n, \mathbb{R})$ -invariant polynomials on $\mathfrak{gl}(n, \mathbb{R})$. So there are corresponding $g'_{2k} \in I^{2k}(\mathrm{GL}(n, \mathbb{R}))$ which restrict to g_{2k} on $\mathfrak{o}(n) \times \cdots \times \mathfrak{o}(n)$. Now a connection ω for the principal bundle $\eta = \mathrm{SO}(\xi)$ extends to a connection ω' for the principal bundle $\eta' = F(\xi)$ of frames of ξ , and Ω' is an extension of Ω . Thus the form

$$g'_{2k}(\Omega') \quad \text{restricts to} \quad g_{2k}(\Omega) \quad \text{on} \quad \mathrm{SO}(E).$$

This shows that $w_{\eta'}(g'_{2k}) = w_\eta(g_{2k}) = p_k(\xi)$. Since $w_{\eta'}(g'_{2k})$ doesn't depend on the particular connection Ω' for $F(\xi)$, we see that we can define $p_k(\xi)$ in terms of an arbitrary connection for $F(\xi)$; it is not necessary to use a connection which preserves inner products, and our bundle does not even have to be orientable.

On the other hand, for $n = 2m$, the Pfaffian $\mathrm{Pf}: \mathfrak{o}(n) \rightarrow \mathbb{R}$ is *not* $\mathrm{GL}(n, \mathbb{R})$ -invariant, nor even $\mathrm{GL}^+(n, \mathbb{R})$ -invariant, so our construction definitely requires orientability, and a connection on $\mathrm{SO}(E)$, i.e., a connection compatible with some metric. Indeed, there are examples (see Milnor and Stasheff [1; pg. 312]) of oriented bundles ξ having a connection ω with $\Omega = 0$, but with $\chi(\xi) \neq 0$; naturally such a connection cannot be compatible with any metric on ξ .

11. COMPLEX BUNDLES

A **complex vector bundle** $\pi: E \rightarrow X$ is defined precisely like a real vector bundle, except that each fibre $\pi^{-1}(x)$ has the structure of a vector space over \mathbb{C} , and in all the conditions for a vector bundle, including local triviality, we replace \mathbb{R} by \mathbb{C} whenever it occurs; vector space isomorphisms are always understood to be isomorphisms of complex vector spaces, hence linear over \mathbb{C} . Linearity over \mathbb{C} is also understood in the definitions of bundle maps (and equivalences) between complex bundles. The Whitney sum $\xi \oplus \eta$ of two complex bundles ξ and η is a complex bundle, and so is the induced bundle $f^*\xi$. The principal bundle $F(\xi)$ of frames is now a principal bundle with group $\mathrm{GL}(n, \mathbb{C}) =$ the set of all non-singular $n \times n$ matrices with complex entries (which may be identified in a natural way with the set of all non-singular linear transformations of \mathbb{C}^n). Note that the Covering Homotopy Theorem (Theorem 4) holds for complex

bundles, since it holds for the corresponding principal bundles. There are two reasons for discussing complex bundles, and their characteristic classes. On the one hand, everything works out to be simpler; on the other hand, there are relations between the characteristic classes for real bundles and those for complex bundles. To discuss complex bundles, however, we need several preliminaries about complex vector spaces.

On the vector space \mathbb{C}^n we could consider the bilinear function

$$(z, w) \mapsto \sum_{i=1}^n z^i w^i.$$

This is not an inner product, since it is not even real, and certainly not positive definite. The linear transformations $T: \mathbb{C}^n \rightarrow \mathbb{C}^n$ which preserve this bilinear function correspond to $n \times n$ complex matrices A such that $AA^t = I$. This group of matrices is known as the **complex orthogonal group**. It is of little interest to us, mainly because it is not compact. We consider instead the function $\mathbb{C}^n \times \mathbb{C}^n \rightarrow \mathbb{C}$ given by

$$\langle z, w \rangle = \sum_{i=1}^n z^i \cdot \overline{w^i}.$$

More generally, for any vector space V over \mathbb{C} we define an **Hermitian inner product** to be a map $\langle \cdot, \cdot \rangle: V \times V \rightarrow \mathbb{C}$ which is linear over \mathbb{C} in the first variable, and which satisfies

$$\begin{aligned} \langle v, w \rangle &= \overline{\langle w, v \rangle} & (\implies \langle v, v \rangle \text{ is real}) \\ \langle v, v \rangle &> 0 & \text{for } v \neq 0. \end{aligned}$$

The first condition shows that $\langle \cdot, \cdot \rangle$ is conjugate linear in the second variable (i.e., $\langle v, w_1 + w_2 \rangle = \langle v, w_1 \rangle + \langle v, w_2 \rangle$ and $\langle v, \alpha w \rangle = \bar{\alpha} \langle v, w \rangle$). Because of the second condition, we can define $|v| = \sqrt{\langle v, v \rangle}$. We compute that

$$\begin{aligned} |v + w|^2 - |v - w|^2 &= 2(\langle v, w \rangle + \overline{\langle v, w \rangle}) \\ \implies |v - iw|^2 - |v + iw|^2 &= 2i(\langle v, w \rangle - \overline{\langle v, w \rangle}). \end{aligned}$$

Consequently, we can express $\langle v, w \rangle$ in terms of $|\cdot|$. A basis v_1, \dots, v_n of V is **orthonormal** with respect to an Hermitian inner product $\langle \cdot, \cdot \rangle$ if we have, precisely as in the real case, $\langle v_i, v_j \rangle = \delta_{ij}$. We can always obtain an orthonormal basis from a given one by the Gram-Schmidt process, which works just as well for Hermitian inner products. Hence, any n -dimensional Hermitian inner product

space $(V, \langle \cdot, \cdot \rangle)$ is isomorphic to \mathbb{C}^n with the standard Hermitian inner product $\langle z, w \rangle = \sum_i z^i \cdot \overline{w^i}$.

We define $U(n) \subset GL(n, \mathbb{C})$ to be the subgroup of all A such that $AA^* = I$, where A^* denotes the conjugate transpose of A ,

$$A^* = \overline{A}^t, \quad \text{i.e.,} \quad A^*_{ij} = \overline{A_{ji}}.$$

We can also think of $U(n)$ as the set of all linear transformations of \mathbb{C}^n which preserve the standard Hermitian inner product. It is easy to see that $U(n)$ is compact, just like $O(n)$. Thus $U(n)$ must be a Lie group (Theorem I.10-15). To see this in a more elementary way, we can consider the exponential map $\exp: (n \times n \text{ complex matrices}) \rightarrow GL(n, \mathbb{C})$ defined, just as in the real case, by

$$\exp M = I + M + \frac{M^2}{2!} + \cdots.$$

Reasoning as on pg. I.388, we easily see that $U(n)$ is a Lie group whose Lie algebra $\mathfrak{u}(n)$ is the set of all $n \times n$ complex matrices M with $M + M^* = 0$ (skew-Hermitian M). Thus $M \in \mathfrak{u}(n)$ if and only if M has the form

$$\begin{pmatrix} ib_{11} & & -B^* \\ & \ddots & \\ B & & ib_{nn} \end{pmatrix} \quad b_{ii} \text{ real.}$$

So $U(n)$ has dimension

$$n + 2(1 + \cdots + n - 1) = n^2.$$

Notice that $U(1)$ is just the set of complex numbers of absolute value 1. Hence $U(1)$ is connected. We can regard $S^{2n-1} \subset \mathbb{C}^n$ as the set of all $z \in \mathbb{C}^n$ with $|z| = 1$. So for $n \geq 2$ we can define $f: U(n) \rightarrow S^{2n-1}$ by $f(A) = A(p_0)$, where p_0 is the n -tuple of complex numbers $(0, \dots, 0, 1)$. Then $f^{-1}(p_0)$ is homeomorphic to $U(n-1)$. Using induction, as in Problem I.3-30, we see that $U(n)$ is connected for all n . (Reasoning similar to that in Problem I.3-31 would show that $GL(n, \mathbb{C})$ is also connected.)

Every vector space V over \mathbb{C} is also a vector space $V_{\mathbb{R}}$ over \mathbb{R} [formally, $V_{\mathbb{R}}$ is V with the same addition map $V \times V \rightarrow V$ and the multiplication $\mathbb{R} \times V \rightarrow V$ which is the restriction of the given multiplication $\mathbb{C} \times V \rightarrow V$]. If v_1, \dots, v_n is a basis for V over \mathbb{C} , then $v_1, iv_1, v_2, iv_2, \dots, v_n, iv_n$ is a basis for $V_{\mathbb{R}}$ over \mathbb{R} . Let $T: V \rightarrow V$ be a linear transformation (over \mathbb{C}) whose matrix with respect to v_1, \dots, v_n is the $n \times n$ complex matrix

$$A = (\alpha_{jk}) = (a_{jk} + ib_{jk}),$$

so that

$$Tv_j = \sum_{k=1}^n \alpha_{kj} v_k.$$

Then the matrix of $T: V_{\mathbb{R}} \rightarrow V_{\mathbb{R}}$ with respect to the basis $v_1, iv_1, \dots, v_n, iv_n$ is the $2n \times 2n$ real matrix $h(A) = (\tilde{\alpha}_{jk})$, where $\tilde{\alpha}_{jk}$ is the 2×2 block

$$\tilde{\alpha}_{jk} = \begin{pmatrix} a_{jk} & -b_{jk} \\ b_{jk} & a_{jk} \end{pmatrix}.$$

It is easy to see, using block multiplication of matrices, that

$$h: \{n \times n \text{ complex matrices}\} \rightarrow \{2n \times 2n \text{ real matrices}\}$$

is a homomorphism. Hence it also gives us a homomorphism

$$h: \text{GL}(n, \mathbb{C}) \rightarrow \text{GL}(2n, \mathbb{R}),$$

and moreover, we easily see that

$$h_* = h: \mathfrak{gl}(n, \mathbb{C}) \rightarrow \mathfrak{gl}(2n, \mathbb{R}).$$

It is also easy to see that

$$h: \text{U}(n) \rightarrow \text{O}(2n).$$

Since $\text{U}(n)$ is connected, and h takes the identity matrix of $\text{U}(n)$ to the identity matrix of $\text{O}(2n)$, we actually have

$$h: \text{U}(n) \rightarrow \text{SO}(2n).$$

This also follows from

48. PROPOSITION. For every $n \times n$ complex matrix A we have

$$\det h(A) = |\det A|^2.$$

PROOF. The formula clearly holds for a diagonal matrix

$$A = \begin{pmatrix} a_{11} + ib_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} + ib_{nn} \end{pmatrix} \Rightarrow h(A) = \begin{pmatrix} a_{11} & -b_{11} & & 0 \\ b_{11} & a_{11} & & \\ & & \ddots & \\ 0 & & & a_{nn} & -b_{nn} \\ & & & b_{nn} & a_{nn} \end{pmatrix}.$$

So it also holds for diagonalizable matrices. But the diagonalizable matrices are dense, and both sides of the equation are continuous in A . So it holds for all A . ♦

If v_1, \dots, v_n and w_1, \dots, w_n are two bases of V and A is the matrix expressing the w 's in terms of the v 's, then $h(A)$ is the matrix expressing the basis $w_1, iw_1, \dots, w_n, iw_n$ in terms of $v_1, iv_1, \dots, v_n, iv_n$. Since $\det h(A) > 0$, this shows that $V_{\mathbb{R}}$ has a natural orientation (which is but a reflection of the fact that $\mathrm{GL}(n, \mathbb{C})$ is connected). If ξ is a complex vector bundle, then we can form a real vector bundle $\xi_{\mathbb{R}}$ by replacing each fibre by the corresponding vector space over \mathbb{R} ; clearly $\xi_{\mathbb{R}}$ is always orientable, with a natural orientation.

For complex vector bundles it is natural to consider **Hermitian metrics**, which assign an Hermitian inner product to each fibre. We can prove they exist, as in the real case, by using partitions of unity (note that a positive real multiple of an Hermitian inner product is also an Hermitian inner product). Using an Hermitian inner product $\langle \cdot, \cdot \rangle$ on the complex bundle $\xi = \pi: E \rightarrow X$ we can define the principal bundle $U(\xi) = \varpi: U(E) \rightarrow X$ with group $U(n)$, whose fibre $\varpi^{-1}(x)$ is the set of all frames of $\pi^{-1}(x)$ which are orthonormal with respect to $\langle \cdot, \cdot \rangle$.

Corresponding to the Grassmannian $G_n(\mathbb{R}^N)$, we have the **complex Grassmannian manifold** $G_n(\mathbb{C}^N)$, consisting of all $W \subset \mathbb{C}^N$ which are subspaces of \mathbb{C}^N (as a vector space over \mathbb{C}) of complex dimension n . If $V_n(\mathbb{C}^N)$ is the set of all linearly independent n -tuples $(v_1, \dots, v_n) \in \mathbb{C}^N \times \dots \times \mathbb{C}^N$, we define the map

$$\rho: V_n(\mathbb{C}^N) \rightarrow G_n(\mathbb{C}^N)$$

by letting

$$\rho((v_1, \dots, v_n)) = (\text{complex}) \text{ subspace of } \mathbb{C}^N \text{ spanned by } v_1, \dots, v_n,$$

and we give $G_n(\mathbb{C}^N)$ the quotient topology for this map. Reasoning exactly as in the real case, we see that $G_n(\mathbb{C}^N)$ can also be described as the left coset space

$$U(N)/U(n) \times U(N-n).$$

Over $G_n(\mathbb{C}^N)$ we have a natural complex bundle $\gamma^n(\mathbb{C}^N)$ defined exactly as in the real case, and for $M > N$ there is a natural map $\alpha: G_n(\mathbb{C}^N) \rightarrow G_n(\mathbb{C}^M)$ such that $\gamma^n(\mathbb{C}^N) \simeq \alpha^* \gamma^n(\mathbb{C}^M)$. The reader may easily check that Theorems 6 and 7 hold for complex bundles when we replace $\gamma^n(\mathbb{R}^N)$ by $\gamma^n(\mathbb{C}^N)$ throughout; the proofs are exactly the same.

To find the characteristic classes for complex bundles, we thus need to compute the cohomology of $U(N)/U(n) \times U(N-n)$. For this we need the solution to two invariance problems. First we want to consider polynomial functions on $\mathbb{C}^n \times \dots \times \mathbb{C}^n$, by which we mean real-valued functions which are polynomials (over \mathbb{R}) in the real and imaginary components of the various vectors.

49. THEOREM. Every polynomial function f of m vectors in \mathbb{C}^n which is invariant under $U(n)$ can be written as a polynomial in the real and imaginary parts of the Hermitian inner products.

Notice that this result is much simpler than the corresponding result for $O(n)$ and $SO(n)$, for there are no determinants involved, even though $U(n)$ is connected. The proof is also simpler, in the sense that various delicate details which arose in the proof of Theorem 35 are not needed. However, certain other considerations are required, and the proof is deferred to Addendum 1.

Another instance of the greater simplicity to be found in the complex domain is afforded by the spectral theorem, which is both more general and easier to prove. We recall that for every linear transformation $T: \mathbb{C}^n \rightarrow \mathbb{C}^n$ there is a unique linear transformation $T^*: \mathbb{C}^n \rightarrow \mathbb{C}^n$, the **adjoint** of T , with

$$\langle Tv, w \rangle = \langle v, T^*w \rangle \quad \text{for } v, w \in \mathbb{C}^n.$$

If T corresponds to the matrix A , then T^* corresponds to the conjugate transpose matrix A^* . We call A **normal** if $AA^* = A^*A$, and similarly for transformations. Both self-adjoint transformations ($T^* = T$) and skew-adjoint transformations ($T^* = -T$) are normal. If T is normal, then

$$\langle Tv, Tv \rangle = \langle v, T^*Tv \rangle = \langle v, TT^*v \rangle = \langle T^*v, T^*v \rangle.$$

Applying this to $T - \lambda I$, which is also normal, we see that

$$|(T - \lambda I)v| = |(T^* - \bar{\lambda} I)v|.$$

Now any linear transformation $T: \mathbb{C}^n \rightarrow \mathbb{C}^n$ has an eigenvector, since the equation $\det(T - \lambda I) = 0$ has a root in the field \mathbb{C} . The above equation shows that if T is normal, then an eigenvector v of T is also an eigenvector of T^* . Therefore the subspace $[v]$ spanned by v is invariant under T^* . Consequently, the orthogonal complement $[v]^\perp$ (under the Hermitian inner product) is invariant under $T^{**} = T$. From the invariance of both $[v]$ and $[v]^\perp$ under T , we easily see, by induction, that T has an orthonormal basis of eigenvectors. Equivalently, for every normal matrix A , there is a matrix $B \in U(n)$ such that BAB^{-1} is a diagonal matrix.

Now it is easy to give a canonical form for elements of $\mathfrak{u}(n)$.

50. PROPOSITION. For every $A \in \mathfrak{u}(n)$ there is a matrix $B \in U(n)$ such that

$$BAB^{-1} = \begin{pmatrix} i\lambda_1 & & 0 \\ & \ddots & \\ 0 & & i\lambda_n \end{pmatrix} \quad \lambda_j \text{ real.}$$

PROOF. Since $A = -A^*$ is normal, there is $B \in U(n)$ such that BAB^{-1} is diagonal. Moreover,

$$(BAB^{-1})^* = B^{-1*}A^*B^* = BA^*B^{-1} = -BAB^{-1},$$

so the diagonal entries of $B^{-1}AB$ must be pure imaginary. ♦

Using this result, we can easily describe the polynomial functions on $\mathfrak{u}(n)$ which are invariant under the adjoint action of $U(n)$. Since the polynomials f_1, \dots, f_n of section 8 are not real-valued on $\mathfrak{u}(n)$, it is convenient to consider instead the polynomials

$$\tilde{f}_k(M) = i^k f_k(M) = f_k(iM),$$

so that

$$\det(I + \lambda iM) = 1 + \lambda \tilde{f}_1(M) + \dots + \lambda^n \tilde{f}_n(M).$$

We also set $f_0(M) = 1$. If $M \in \mathfrak{u}(n)$, then for all real λ we have

$$(I + \lambda iM)^* = I + \lambda iM \implies \overline{\det(I + \lambda iM)} = \det(I + \lambda iM),$$

which shows that all $\tilde{f}_i(M)$ are real. It is easy to see, as on page 337, that for all $A \in \mathfrak{gl}(r, \mathbb{C})$ and $B \in \mathfrak{gl}(s, \mathbb{C})$ we have

$$\tilde{f}_k(A \oplus B) = \sum_{l=0}^k \tilde{f}_l(A) \cdot \tilde{f}_{k-l}(B).$$

51. THEOREM. Every polynomial function f on $\mathfrak{u}(n)$ which is invariant under the adjoint action of $U(n)$ is a polynomial in $\tilde{f}_1, \dots, \tilde{f}_n$.

PROOF. For $\lambda_1, \dots, \lambda_n \in \mathbb{C}$, let $[\lambda_1, \dots, \lambda_n]$ be the diagonal matrix with entries $i\lambda_1, \dots, i\lambda_n$ on the diagonal. Define

$$g(\lambda_1, \dots, \lambda_n) = f([\lambda_1, \dots, \lambda_n]).$$

Then g is a polynomial in $\lambda_1, \dots, \lambda_n$. Moreover, g is symmetric, since

$$[\lambda_{\pi(1)}, \dots, \lambda_{\pi(n)}] = A \cdot [\lambda_1, \dots, \lambda_n] A^{-1}$$

where $A \in U(n)$ is a suitable permutation matrix. So we can write

$$g(\lambda_1, \dots, \lambda_n) = p(\sigma_1(i\lambda_1, \dots, i\lambda_n), \dots, \sigma_n(i\lambda_1, \dots, i\lambda_n))$$

for some polynomial p . Then

$$f([\lambda_1, \dots, \lambda_n]) = p(\tilde{f}_1([\lambda_1, \dots, \lambda_n]), \dots, \tilde{f}_n([\lambda_1, \dots, \lambda_n])).$$

The result follows as before, using Proposition 50. ♦

Using Theorems 49 and 51, we can carry out the whole program of section 9 for

$$G_n(\mathbb{C}^N) = \mathrm{U}(N)/\mathrm{U}(n) \times \mathrm{U}(N-n) = \mathrm{U}(N)/H.$$

A bi-invariant Riemannian metric $\langle \cdot, \cdot \rangle$ on $\mathrm{U}(n)$ can be defined explicitly as follows. For $M, P \in \mathfrak{u}(n)$, let

$$\langle M, P \rangle = \mathrm{Re}(\mathrm{trace} \, MP^*) = \mathrm{Re} \sum_{i,j} M_{ij} \cdot \overline{P_{ij}}.$$

As in the case of $\mathrm{O}(n)$, if we extend $\langle \cdot, \cdot \rangle$ to $\mathrm{U}(n)$ by left invariance, then it will also be right invariant. Now \mathfrak{h}^\perp consists of all matrices

$$\begin{pmatrix} 0 & P \\ -P^* & 0 \end{pmatrix} \quad P \text{ an } n \times (N-n) \text{ complex matrix.}$$

On $\mathfrak{h}^\perp \times \mathfrak{h}^\perp$ we have the functions

$$(P, Q) \mapsto \sum_r P_i^r \cdot \overline{Q_j^r}$$

(which are bilinear *over* \mathbb{R}); their alternations ψ_{ij} , given by

$$\psi_{ij}(P, Q) = \sum_r P_i^r \cdot \overline{Q_j^r} - Q_i^r \cdot \overline{P_j^r},$$

are complex-valued alternating bilinear functions on \mathfrak{h}^\perp . A form $\eta \in \Omega^k(\mathfrak{h}^\perp)$ will be invariant under $\{I\} \times \mathrm{U}(N-n)$ if and only if it is a linear combination of wedge products of the forms $\mathrm{Re} \, \psi_{ij}$ and $\mathrm{Im} \, \psi_{ij}$ (since there are no determinants to worry about in the first main theorem of invariance theory for $\mathrm{U}(n)$, we do not need $k < N-n$). Such linear combinations can be described as $f(\psi)$ for polynomial functions $f: \mathfrak{u}(n) \rightarrow \mathbb{R}$ [this representation is unique in dimensions $< N-n$], and the combinations which are invariant under $\mathrm{U}(n) \times \{I\}$ correspond to functions $f: \mathfrak{u}(n) \rightarrow \mathbb{R}$ which are invariant under the adjoint action of $\mathrm{U}(n)$. Thus we have a class $\tilde{f}_i(\psi) \in H^{2i}(G_n(\mathbb{C}^N))$ for each of the polynomials $\tilde{f}_1, \dots, \tilde{f}_n$ of Theorem 51, and every element of $H^*(\mathrm{U}(N)/\mathrm{U}(n) \times \mathrm{U}(N-n))$ is a linear combination of cup products of these elements [it is a *unique* linear combination in dimensions $< N-n$].

The analogue of Proposition 41 holds, so we will consider only N with $2n < N-n$. Then all elements of $H^*(G_n(\mathbb{C}^N))$ in dimensions $< N-n$ are unique linear combinations of cup products of the classes corresponding to

$$\tilde{f}_1(\psi), \dots, \tilde{f}_n(\psi).$$

We let

$$c_{n;k} \in H^{2k}(G_n(\mathbb{C}^N)) \quad k = 1, \dots, n$$

be the class corresponding to

$$\frac{1}{(2\pi)^k} \tilde{f}_k(\psi).$$

For an n -dimensional complex bundle ξ over M we define

$$c_k(\xi) = g^* c_{n;k},$$

where

$$g: M \rightarrow G_n(\mathbb{C}^N) \quad \text{satisfies} \quad g^* \gamma^n(\mathbb{C}^N) \simeq \xi.$$

The characteristic class $\xi \mapsto c_k(\xi)$ is called the k^{th} **Chern class** for n -dimensional complex bundles; every characteristic class for n -dimensional complex bundles is a polynomial in the Chern classes c_1, \dots, c_n . The class

$$c(\xi) = 1 + c_1(\xi) + \dots + c_n(\xi) = c_0(\xi) + c_1(\xi) + \dots + c_n(\xi)$$

is called the **total Chern class** of the n -dimensional complex bundle ξ .

Just as in the real case, the Chern classes of an n -dimensional complex bundle $\xi = \pi: E \rightarrow M$ may be described in terms of a connection. Choose any Hermitian metric for ξ , form the corresponding principal $U(n)$ bundle $U(\xi) = \varpi: U(E) \rightarrow M$, and let ω be a $\mathfrak{u}(n)$ -valued connection on $U(\xi)$, with $\mathfrak{u}(n)$ -valued curvature form Ω .

52. THEOREM. The k^{th} Chern class $c_k(\xi)$ of ξ is represented by the unique form Λ on M with

$$\varpi^* \Lambda = \frac{1}{(2\pi)^k} \tilde{f}_k(\Omega).$$

In other words, we have

$$c_k(\xi) = w(g_k),$$

where $g_k = \mathcal{J}(\tilde{f}_k) \in I^{2k}(U(n))$ corresponds to the polynomial function \tilde{f}_k on $\mathfrak{u}(n)$, and w is the Weil homomorphism for $U(\xi)$.

PROOF. First of all, an obvious analogue of Corollary 5 shows that the principal bundles η are all equivalent, no matter what Hermitian metric we choose. Now to prove the result, we just have to consider the universal bundle $\gamma^n(\mathbb{C}^N)$. This has a natural Hermitian metric, just like the natural Riemannian metric for $\tilde{\gamma}^n(\mathbb{R}^N)$, on page 345. All the succeeding considerations also have natural analogues, and the result follows exactly as in the proof of Theorem 42 (or Theorem 43). ♦

As an immediate consequence we have

53. THEOREM. If ξ and η are complex bundles over M , then the total Chern class of $\xi \oplus \eta$ is given by the **Whitney product formula**

$$c(\xi \oplus \eta) = c(\xi) \cup c(\eta).$$

PROOF. Exactly like the proof of Theorem 45, except now using the formula on page 362. ♦

We can also find relationships between Chern classes and Pontryagin classes. For a real vector space V , we define a complex vector space $V_{\mathbb{C}}$ by letting $V_{\mathbb{C}} = V \oplus V$, with complex multiplication determined by

$$i \cdot (v, w) = (-w, v) \quad [\text{thus we think of } (v, w) \text{ as } v + iw].$$

Doing this in each fibre of a real vector bundle ξ gives a complex vector bundle $\xi_{\mathbb{C}}$.

54. THEOREM. If $\xi = \pi: E \rightarrow M$ is an oriented n -dimensional vector bundle, then

$$c_{2k}(\xi_{\mathbb{C}}) = (-1)^k p_k(\xi) \quad k = 1, \dots, [n/2].$$

PROOF. Choose the Hermitian metric on $\xi_{\mathbb{C}} = \pi': E_{\mathbb{C}} \rightarrow M$ to be an extension of a Riemannian metric on ξ . Then $\text{SO}(E) \subset \text{U}(E_{\mathbb{C}})$, and the projection $\varpi: \text{SO}(E) \rightarrow M$ is the restriction of the projection $\varpi': \text{U}(E_{\mathbb{C}}) \rightarrow M$. A connection ω on $\text{SO}(E)$ has a unique extension to a $(\mathfrak{u}(n))$ -valued connection ψ on $\text{U}(E_{\mathbb{C}})$ [as in the proof of Theorem 22, ψ is determined on the new vertical vectors of $\text{SO}(E)$, and hence on all of $\text{U}(E_{\mathbb{C}})$]. Let Ψ be the curvature form of ψ . At any point $e \in \text{SO}(E)$ the horizontal vectors for ψ are the same as for ω , so at e we have

$$\begin{aligned} \Psi &= \Omega \quad [\text{on tangent vectors to } \text{SO}(E)] \\ \implies \tilde{f}_{2k}(\Psi) &= \tilde{f}_{2k}(\Omega) = (-1)^k f_{2k}(\Omega). \end{aligned}$$

So if Λ represents $p_k(\xi)$ and Υ represents $c_{2k}(\xi_{\mathbb{C}})$, then at e we have

$$\begin{aligned} \varpi^* \Lambda &= \frac{1}{(2\pi)^k} f_{2k}(\Omega) \\ &= (-1)^k \frac{1}{(2\pi)^k} \tilde{f}_{2k}(\Psi) \\ &= (-1)^k \varpi'^* \Upsilon. \end{aligned}$$

This implies that

$$\Lambda = (-1)^k \Upsilon. \quad \spadesuit$$

The odd Chern classes, which are missing in Theorem 54, are easily calculated by the following considerations. Every complex vector space V gives rise to another complex vector space \bar{V} in which complex multiplication \bullet is defined by

$$\alpha \bullet v = \bar{\alpha} \cdot v.$$

Applying this process to each fibre of a complex bundle ξ we get the **conjugate bundle** $\bar{\xi}$. The bundles ξ and $\bar{\xi}$ are equivalent as real bundles, of course, but there may not be an equivalence which is complex linear on each fibre.

55. PROPOSITION. If ξ is a complex vector bundle, then

$$c(\bar{\xi}) = 1 - c_1(\xi) + c_2(\xi) - c_3(\xi) + \cdots.$$

PROOF. If $\varpi: U(E) \rightarrow M$ is the associated principal bundle $U(\xi)$ for ξ , with R_A the right multiplication by $A \in U(n)$, then for the principal bundle $U(\bar{\xi})$ we may choose the same total space $U(E)$, but with the action \bar{R} of $U(n)$ on the right given by

$$\bar{R}_A = R_{\bar{A}}.$$

So if ω is a connection on $U(\xi)$, with curvature form Ω , then $\bar{\omega}$ (the complex conjugate of ω) will be a connection on $U(\bar{\xi})$, with curvature form $\bar{\Omega}$. If Λ represents $c_k(\xi)$ and Υ represents $c_k(\bar{\xi})$, then

$$\begin{aligned} \varpi^* \Upsilon &= \frac{1}{(2\pi)^k} \tilde{f}_k(\bar{\Omega}) \\ &= \frac{1}{(2\pi)^k} \tilde{f}_k(-\Omega^t) \\ &= \frac{(-1)^k}{(2\pi)^k} \tilde{f}_k(\Omega) \\ &= (-1)^k \varpi^* \Lambda. \end{aligned}$$

Hence $\Upsilon = (-1)^k \Lambda. \quad \spadesuit$

56. COROLLARY. If ξ is a real vector bundle, then

$$c_k(\xi_{\mathbb{C}}) = 0 \quad \text{for } k \text{ odd.}$$

PROOF. We just have to note that

$$\xi_{\mathbb{C}} \simeq \bar{\xi}_{\mathbb{C}} \quad (\text{as complex bundles}).$$

This is due to the fact that there is a natural complex isomorphism $V_{\mathbb{C}} \rightarrow \bar{V}_{\mathbb{C}}$ for every real vector space V —we merely take $(v, w) \mapsto (v, -w)$. ♦

Instead of starting with a real bundle, we can instead begin with a complex bundle ξ , and regard it as an oriented real bundle $\xi_{\mathbb{R}}$ (of even dimension). In order to find its Pontryagin and Euler classes, we need a lemma concerning the homomorphism $h: \mathrm{GL}(n, \mathbb{C}) \rightarrow \mathrm{GL}(2n, \mathbb{R})$ defined on page 359. Since $h = h_*: \mathfrak{gl}(n, \mathbb{C}) \rightarrow \mathfrak{gl}(2n, \mathbb{R})$, we have $h(\mathfrak{u}(n)) \subset \mathfrak{o}(2n)$.

57. LEMMA. For $M \in \mathfrak{u}(n)$ we have

$$f_{2k}(h(M)) = (-1)^k \sum_{l=0}^{2k} (-1)^l \tilde{f}_l(M) \tilde{f}_{2k-l}(M),$$

$$\mathrm{Pf}(h(M)) = \tilde{f}_n(M).$$

PROOF. For all real λ we have

$$\begin{aligned} 1 + \lambda^2 f_2(h(M)) + \cdots + \lambda^{2n} f_{2n}(h(M)) &= \det(I_{2n} + \lambda h(M)) & I_{2n} &= \text{identity of } \mathrm{O}(2n) \\ &= \det h(I_n + \lambda M) & I_n &= \text{identity of } \mathrm{U}(n) \\ &= |\det(I_n + \lambda M)|^2 & & \text{by Proposition 48} \\ &= |1 - i\lambda \tilde{f}_1(M) - \lambda^2 \tilde{f}_2(M) + i\lambda^3 \tilde{f}_3(M) + \lambda^4 \tilde{f}_4(M) - \cdots|^2 \\ &= |(1 - \lambda^2 \tilde{f}_2(M) + \lambda^4 \tilde{f}_4(M) - \cdots) - i(\lambda \tilde{f}_1(M) - \lambda^3 \tilde{f}_3(M) + \cdots)|^2 \\ &= (1 - \lambda^2 \tilde{f}_2(M) + \lambda^4 \tilde{f}_4(M) - \cdots)^2 + (\lambda \tilde{f}_1(M) - \lambda^3 \tilde{f}_3(M) + \cdots)^2. \end{aligned}$$

The coefficient of λ^{2k} on the right side is

$$\begin{aligned} &\sum_{l \text{ even}} (-1)^{\frac{l}{2}} (-1)^{\frac{2k-l}{2}} \tilde{f}_l(M) \tilde{f}_{2k-l}(M) \\ &\quad + \sum_{l \text{ odd}} (-1)^{\frac{l-1}{2}} (-1)^{\frac{2k-l-1}{2}} \tilde{f}_l(M) \tilde{f}_{2k-l}(M) \\ &= (-1)^k \left[\sum_{l \text{ even}} \tilde{f}_l(M) \tilde{f}_{2k-l}(M) - \sum_{l \text{ odd}} \tilde{f}_l(M) \tilde{f}_{2k-l}(M) \right] \\ &= (-1)^k \sum_l (-1)^l \tilde{f}_l(M) \tilde{f}_{2k-l}(M). \end{aligned}$$

For the Pfaffian we have

$$\begin{aligned} [\text{Pf}(h(M))]^2 &= \det h(M) = |\det M|^2 = |f_n(M)|^2 \\ &= \left| \frac{\tilde{f}_n(M)}{i^n} \right|^2 = |\tilde{f}_n(M)|^2, \end{aligned}$$

and hence

$$\text{Pf}(h(M)) = \pm \tilde{f}_n(M).$$

To settle the sign, we consider

$$M = \begin{pmatrix} i & & 0 \\ & \ddots & \\ 0 & & i \end{pmatrix} \Rightarrow h(M) = \begin{pmatrix} -S & & 0 \\ & \ddots & \\ 0 & & -S \end{pmatrix} \quad S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then

$$\text{Pf}(h(M)) = (-1)^n = \det iM = \tilde{f}_n(M),$$

so the + sign is correct. ♦

The relevance of this Lemma will become immediately apparent in the proof of our final result.

58. THEOREM. If ξ is a complex bundle of dimension n , then

$$p_k(\xi_{\mathbb{R}}) = (-1)^k \sum_{l=0}^{2k} (-1)^l c_l(\xi) \cup c_{2k-l}(\xi) \quad k = 1, \dots, 2n,$$

and

$$\chi(\xi_{\mathbb{R}}) = c_n(\xi).$$

PROOF. Note that an Hermitian inner product $\langle \cdot, \cdot \rangle$ on a complex vector space V gives an ordinary inner product on $V_{\mathbb{R}}$ —we define $v_1, iv_1, \dots, v_n, iv_n$ to be orthonormal whenever v_1, \dots, v_n is orthonormal with respect to $\langle \cdot, \cdot \rangle$. This inner product on $V_{\mathbb{R}}$ is well-defined, for if $w_j = \sum_l a_{lj} v_l$ is another orthonormal basis, then the matrix $A = (a_{lj})$ is clearly in $U(n)$, so $h(A) \in SO(n)$. Choosing an Hermitian metric on $\xi = \pi: E \rightarrow M$, and applying this construction to each fibre, we obtain a Riemannian metric on $\xi_{\mathbb{R}}$. Moreover, for the corresponding principal bundles, we have $U(E) \subset SO(E)$, and the projection $\varpi: U(E) \rightarrow M$ is the restriction of the projection $\varpi': SO(E) \rightarrow M$.

For $A \in U(n)$, the right action R_A on $U(E)$ is just the restriction of the right action $R_{h(A)}$ on $SO(E)$. A connection ω on $U(E)$ is $u(n)$ -valued, as is its curvature form Ω . The $h(u(n))$ -valued form $h \circ \omega$ has a unique extension to a connection ψ on $SO(E)$, with curvature form Ψ . At any point $e \in U(E)$ we have

$$\Psi = h \circ \Omega \quad \text{on tangent vectors to } U(E).$$

So if Λ represents $p_k(\xi_{\mathbb{R}})$ and Λ_l represents $c_l(\xi)$ for $l = 0, \dots, 2k$, then at e we have

$$\begin{aligned} \varpi'^* \Lambda &= \frac{1}{(2\pi)^{2k}} f_{2k}(\Psi) \\ &= \frac{1}{(2\pi)^{2k}} f_{2k}(h \circ \Omega) \\ &= \frac{1}{(2\pi)^{2k}} (-1)^k \sum_{l=0}^{2k} (-1)^l \tilde{f}_l(\Omega) \wedge \tilde{f}_{2k-l}(\Omega) \quad \text{by Lemma 57} \\ &= (-1)^k \sum_{l=0}^{2k} (-1)^l \varpi^* \Lambda_l \wedge \varpi^* \Lambda_{2k-l}, \end{aligned}$$

which proves the first formula.

If Λ represents $\chi(\xi_{\mathbb{R}})$, then at e we have

$$\begin{aligned} \varpi'^* \Lambda &= \frac{1}{(2\pi)^n} \text{Pf}(\Psi) \\ &= \frac{1}{(2\pi)^n} \text{Pf}(h \circ \Omega) \\ &= \frac{1}{(2\pi)^n} \tilde{f}_n(\Omega) \quad \text{by Lemma 57} \\ &= \varpi^* \Lambda_n. \end{aligned}$$

This proves the second formula. ♦

12. VALEDICTORY

Now that we have built up so much machinery, it seems a shame not to use it. But this would really take us out of the field of differential geometry entirely. We have tried to show how the characteristic classes arise naturally, how they can be computed by differential geometric means, why they should be

expressible in terms of curvature, and especially how the Euler class is expressed in terms of $K_n dV$, which involves $\text{Pf}(\Omega)$. For further applications of these characteristic classes, the reader is urged to consult books specifically devoted to the subject, where characteristic classes are usually defined by methods of algebraic topology. One of the most famous set of notes, now finally available in book form, is Milnor and Stasheff [1]. Here the Euler class is defined essentially as we have defined it, in terms of the Thom class. But the Pontryagin and Chern classes are defined in completely different ways. For a complex n -dimensional bundle $\xi = \pi: E \rightarrow X$, the top Chern class $c_n(\xi)$ is *defined* by the formula of Theorem 58, as

$$c_n(\xi) = \chi(\xi_{\mathbb{R}}).$$

The other Chern classes are defined inductively, as follows. Choosing an Hermitian metric for ξ , we form the associated sphere bundle $S \subset E$, and let $\pi_0: S \rightarrow X$ be the restriction of π . It turns out that the map

$$\pi_0^*: H^k(X) \rightarrow H^k(S)$$

is an isomorphism for $k < n$. Now $\pi_0^*\xi$ has a section, so it can be written as the Whitney sum

$$\pi_0^*\xi = \xi_1 \oplus \xi_2,$$

where ξ_2 is a trivial 1-dimensional complex bundle. The Chern class $c_{n-1}(\xi)$ is defined as

$$c_{n-1}(\xi) = (\pi_0^*)^{-1}(c_{n-1}(\xi_1)).$$

This makes sense, since c_{n-1} is defined for the $(n-1)$ -dimensional complex bundle ξ_1 . Moreover, it is compatible with our definition, since it is equivalent to

$$c_{n-1}(\pi_0^*\xi) = c_{n-1}(\xi_1),$$

which is what the Whitney product formula gives when ξ_2 is trivial. Now that $c_{n-1}(\xi)$ is defined for all n -dimensional complex bundles ξ , the Chern class $c_{n-2}(\xi)$ can be defined as

$$c_{n-2}(\xi) = (\pi_0^*)^{-1}(c_{n-1}(\xi_1)),$$

and so on, by induction. After the Chern classes are defined, the Whitney product formula is proved, and the cohomology of $G_n(\mathbb{C}^\infty)$ is calculated, by means of various tricks. The Pontryagin classes are defined by the formula in Theorem 54, and all properties are derived from this definition and the properties of the Chern classes. In the whole development, there is no restriction

to manifolds, and singular homology is used throughout. Moreover, the coefficients are \mathbb{Z} , rather than \mathbb{R} . Integer coefficients can be used because the Thom class can be defined with integer coefficients, so the Euler class has integer coefficients, and the map $\pi_0^*: H^k(X) \rightarrow H^k(S)$ is an isomorphism with integer coefficients. This shows, by the way, that our Euler, Pontryagin, and Chern classes, defined originally with real coefficients, are actually all integral classes; that, of course, was the reason for inserting the various factors $(2\pi)^{-\alpha}$ in the definitions.

ADDENDUM 1

INVARIANT THEORY FOR THE UNITARY GROUP

At first sight, the problem of determining all invariant polynomials for $U(n)$ seems peculiarly complicated, since we are dealing with polynomial functions of the real and imaginary parts of the components of the vectors, rather than polynomial functions of the components themselves. But there is a trick which will allow us to reduce the problem to one where we study only polynomials of the latter type. The basic result which we need for such polynomials can actually be formulated for any field.

Let k be an arbitrary infinite field, and let $V = k^n$ be the standard n -dimensional vector space over k , with standard basis elements e_1, \dots, e_n . The group of all non-singular $n \times n$ matrices with entries in k is denoted by $GL(n, k)$. Just as in the real case, a matrix $A = (a_{ij}) \in GL(n, k)$ is also regarded as a linear transformation $A: V \rightarrow V$, by the rule

$$A(e_i) = \sum_{j=1}^n a_{ji} e_j,$$

so that for a (row vector) $v \in k^n$ we have

$$A(v) = v \cdot A^t.$$

We also define an action of $GL(n, k)$ on the m -fold product $V \times \dots \times V$ by

$$A \cdot (v_1, \dots, v_m) = (A(v_1), \dots, A(v_m)).$$

A function

$$f: V \times \dots \times V \rightarrow k$$

is called a polynomial function if it is a polynomial (over k) in the components of the elements of V ; we also define homogeneous polynomial functions just as before. We say that f is invariant under a group $G \subset GL(n, k)$ if

$$f(A \cdot v) = f(v) \quad \text{for all } v \in V \times \dots \times V, \text{ and all } A \in G.$$

We still have the standard basis vectors e_{ri} for $V \times \dots \times V$, and the partial derivatives $\partial f / \partial e_{ri}$ can be defined formally for polynomial functions f . Euler's theorem can be checked formally, polarizations are defined as before, and the Capelli identities still hold. Introducing the partial ordering on the polynomial functions as before, we still have assertion (A) on page 327.

Actually, we want to be even more general, and consider functions

$$f: \underbrace{V \times \cdots \times V}_m \times \underbrace{V^* \times \cdots \times V^*}_l \rightarrow k.$$

We define an action of $\mathrm{GL}(n, k)$ on $V \times \cdots \times V^*$ by

$$A \cdot (v_1, \dots, v_m, \phi_1, \dots, \phi_l) = (A(v_1), \dots, A(v_m), \phi_1 \circ A^{-1}, \dots, \phi_l \circ A^{-1}),$$

and we say that f is invariant under a group $G \subset \mathrm{GL}(n, k)$ if

$$f(A \cdot (v, \phi)) = f((v, \phi)) \quad \text{for all } (v, \phi) \in V \times \cdots \times V^*, \text{ and } A \in G.$$

For example, the “evaluations”

$$\varepsilon_{rs}(v_1, \dots, v_m, \phi_1, \dots, \phi_l) = \phi_s(v_r)$$

are invariant under all of $\mathrm{GL}(n, k)$. It will be convenient to identify an element $\phi \in V^*$ with the *column* vector

$$\xi = \begin{pmatrix} \phi(e_1) \\ \vdots \\ \phi(e_n) \end{pmatrix}.$$

Then it turns out that the action of $\mathrm{GL}(n, k)$ is

$$A \cdot (v_1, \dots, v_m, \xi_1, \dots, \xi_l) = (v_1 \cdot A^t, \dots, v_m \cdot A^t, (A^{-1})^t \cdot \xi_1, \dots, (A^{-1})^t \cdot \xi_l).$$

We call f a polynomial function if $f(v_1, \dots, v_m, \xi_1, \dots, \xi_l)$ is a polynomial in the v_{ri} and ξ_{sj} (here ξ_{sj} is the entry in the j^{th} row of ξ_s). Notice that the evaluations are polynomial functions—under the identification of V^* with the set of column vectors they are simply the maps

$$\varepsilon_{rs}(v_1, \dots, v_m, \xi_1, \dots, \xi_l) = v_r \cdot \xi_s.$$

Two other important types of polynomial functions are the functions

$$\det_{r_1, \dots, r_n}(v_1, \dots, v_m, \xi_1, \dots, \xi_l) = \det \begin{pmatrix} v_{r_1} \\ \vdots \\ v_{r_n} \end{pmatrix}$$

and

$$\det_{s_1, \dots, s_n}^*(v_1, \dots, v_m, \xi_1, \dots, \xi_l) = \det(\xi_{s_1}, \dots, \xi_{s_n}).$$

They are invariant under the subgroup $\mathrm{SL}(n, k) \subset \mathrm{GL}(n, k)$ consisting of matrices of determinant 1.

We can define homogeneous polynomial functions as before, except now we must consider functions which are homogeneous of degree $(\alpha_1, \dots, \alpha_m)$ in the V variables, and of degree $(\beta_1, \dots, \beta_l)$ in the V^* variables. For any homogeneous f , we can apply the apparatus of the Capelli identities to either the V variables or the V^* variables separately, and obtain assertion (A) on page 327 where the partial ordering $<$ is applied to either the degree in V or the degree in V^* , the polarizations being applied to the variables in V or V^* , respectively.

59. THEOREM. For all m, l , and n we have

$SL_n^{m,l}$: Every polynomial function f of m vectors in k^n and l vectors in $(k^n)^*$ which is invariant under $SL(n, k)$ can be written as a polynomial in the evaluation functions ε_{rs} and the determinant functions \det_{r_1, \dots, r_n} and \det_{s_1, \dots, s_n}^* .

PROOF. The proof is similar to that of Theorem 35, but the steps are easier. First we note that $SL_n^{m,l} \implies SL_n^{m',l'}$ for $m' \leq m$ and $l' \leq l$. Next we have

60. LEMMA. If $SL_n^{n-1, n-1}$ holds, then $SL_n^{m,l}$ holds for all $m \geq n-1, l \geq n-1$.

PROOF. The argument is similar to that for Lemma 36. In the present case we can start with $SL_n^{n-1, n-1}$ rather than $SL_n^{n,n}$ because the term \det appearing in assertion (A) in the case $m = n$ causes no problems—it is one of the invariants in terms of which we are trying to express f . **Q.E.D.**

Now the proof of the special case to which we have reduced the problem does not even require an inductive argument:

61. LEMMA. $SL_n^{n-1, n-1}$ holds for all n .

PROOF. Consider $(v_1, \dots, v_{n-1}, \phi_1, \dots, \phi_{n-1})$ satisfying

$$(*) \quad \det(\phi_i(v_j)) \neq 0.$$

Then $\bigcap_{i=1}^{n-1} \ker \phi_i$ is 1-dimensional. Let $0 \neq w \in \bigcap_i \ker \phi_i$. If we had

$$w = \sum_{j=1}^{n-1} a_j v_j$$

for constants a_i (necessarily not all zero), then we would have

$$0 = \phi_i(w) = \sum_{j=1}^{n-1} a_j \phi_i(v_j) \quad j = 1, \dots, n-1,$$

contradicting (*). So w is linearly independent of v_1, \dots, v_{n-1} . Hence there is some $A \in \text{SL}(n, k)$ such that

$$\begin{aligned} A(e_i) &= v_i \quad i = 1, \dots, n-1 \\ A(e_n) &= \text{a multiple of } w. \end{aligned}$$

Then

$$\begin{aligned} (1) \quad & f(v_1, \dots, v_{n-1}, \phi_1, \dots, \phi_{n-1}) \\ &= f(A^{-1}(v_1), \dots, A^{-1}(v_{n-1}), \phi_1 \circ A, \dots, \phi_{n-1} \circ A) \\ &= f(e_1, \dots, e_{n-1}, \phi_1 \circ A, \dots, \phi_{n-1} \circ A); \end{aligned}$$

note that

$$\begin{aligned} (2) \quad & (\phi_i \circ A)(e_j) = \phi_i(v_j) \quad j = 1, \dots, n-1 \\ & (\phi_i \circ A)(e_n) = 0. \end{aligned}$$

Now define a polynomial function F of $(n-1)^2$ variables a_{ij} , as follows:

$$\begin{aligned} F(\{a_{ij}\}) &= f(e_1, \dots, e_{n-1}, \mu_1, \dots, \mu_{n-1}), \\ &\text{where } \mu_i \text{ are the unique linear functionals with} \\ &\mu_i(e_j) = a_{ij} \quad j = 1, \dots, n-1 \\ &\mu_i(e_n) = 0. \end{aligned}$$

Then equations (1) and (2) show that

$$(**) \quad f(v_1, \dots, v_{n-1}, \phi_1, \dots, \phi_{n-1}) = F(\{\phi_i(v_j)\})$$

whenever (*) holds. A standard argument ("the principle of irrelevance of algebraic inequalities") shows that consequently (**) holds everywhere: for the polynomial

$$[f(v_1, \dots, v_{n-1}, \phi_1, \dots, \phi_{n-1}) - F(\{\phi_i(v_j)\})] \cdot \det(\phi_i(v_j))$$

is identically 0, hence one of the factors must be identically 0, and the second factor certainly isn't. ♦

We will use Theorem 59 only for the case $\mathrm{SL}(n, \mathbb{C})$. We easily see (compare Problem I.10-27) that the Lie algebra $\mathfrak{sl}(n, \mathbb{C})$ of $\mathrm{SL}(n, \mathbb{C})$ consists of all $n \times n$ complex matrices with trace $= 0$. Similarly (compare page 358), the group $\mathrm{SU}(n) = \mathrm{U}(n) \cap \mathrm{SL}(n, \mathbb{C})$ has Lie algebra $\mathfrak{su}(n)$ consisting of all matrices

$$\begin{pmatrix} ib_{11} & & -B^* \\ & \ddots & \\ B & & ib_{nn} \end{pmatrix} \quad \begin{array}{l} b_{ii} \text{ real} \\ \sum b_{ii} = 0. \end{array}$$

Notice that $\mathfrak{su}(n)$ is not a complex subspace of $\mathfrak{gl}(n, \mathbb{C})$; however, it is easy to find the complex subspace $W \subset \mathfrak{gl}(n, \mathbb{C})$ spanned by $\mathfrak{su}(n)$. Note that W must contain

$$-i \cdot \begin{pmatrix} ib_{11} & & 0 \\ & \ddots & \\ 0 & & ib_{nn} \end{pmatrix} = \begin{pmatrix} b_{11} & & 0 \\ & \ddots & \\ 0 & & b_{nn} \end{pmatrix} \quad \begin{array}{l} b_{ii} \text{ real} \\ \sum b_{ii} = 0 \end{array}$$

and

$$-i \cdot \begin{pmatrix} 0 & -(iA)^* \\ & \ddots \\ iA & & 0 \end{pmatrix} = -i \cdot \begin{pmatrix} 0 & +iA^* \\ & \ddots \\ iA & & 0 \end{pmatrix} = \begin{pmatrix} 0 & & A^* \\ & \ddots & \\ A & & 0 \end{pmatrix},$$

and thus also the matrix

$$\begin{pmatrix} 0 & & 0 \\ & \ddots & \\ A & & 0 \end{pmatrix} = \begin{pmatrix} 0 & & -A^* \\ & \ddots & \\ A & & 0 \end{pmatrix} + \begin{pmatrix} 0 & & A^* \\ & \ddots & \\ A & & 0 \end{pmatrix}, \quad \text{as well as} \quad \begin{pmatrix} 0 & & A \\ & \ddots & \\ 0 & & 0 \end{pmatrix}.$$

From this we easily see that

(*) the complex subspace of $\mathfrak{gl}(n, \mathbb{C})$ spanned by $\mathfrak{su}(n)$ is just $\mathfrak{sl}(n, \mathbb{C})$.

This simple fact leads to

62. LEMMA. Let $g: \mathrm{GL}(n, \mathbb{C}) \rightarrow \mathbb{C}$ be a complex analytic function (this makes sense, since $\mathrm{GL}(n, \mathbb{C})$ is an open subset of \mathbb{C}^{n^2}). If g vanishes on $\mathrm{SU}(n)$, then g also vanishes on $\mathrm{SL}(n, \mathbb{C})$.

PROOF. Pick $Y_1, Y_2 \in \mathfrak{su}(n)$, and consider the function $h: \mathbb{C} \rightarrow \mathbb{C}$ defined by

$$h(z) = g(\exp(zY_1 + Y_2)).$$

Then h is analytic and vanishes for all real z . So h vanishes for all z . Similarly, we may now prove that

$$g(\exp(z_1 Y_1 + z_2 Y_2)) = 0 \quad \text{for all } z_1, z_2 \in \mathbb{C}.$$

Then (*) implies that $g(\exp(X)) = 0$ for all $X \in \mathfrak{sl}(n, \mathbb{C})$. But the image of $\exp: \mathfrak{sl}(n, \mathbb{C}) \rightarrow \mathrm{SL}(n, \mathbb{C})$ is dense* in $\mathrm{SL}(n, \mathbb{C})$, since the diagonalizable matrices are certainly in the image of \exp . Hence $g = 0$ on all of $\mathrm{SL}(n, \mathbb{C})$. ♦

63. COROLLARY. Let f be a polynomial function of m vectors of \mathbb{C}^n and l vectors of $(\mathbb{C}^n)^*$ [that is, f is a polynomial over \mathbb{C} in the (complex) components of the vectors]. If f is invariant under $\mathrm{SU}(n)$, then f is also invariant under $\mathrm{SL}(n, \mathbb{C})$, and is thus a polynomial in the evaluation functions ε_{rs} and the determinant functions \det_{r_1, \dots, r_n} and \det_{s_1, \dots, s_n}^* .

PROOF. For fixed $(v, \phi) \in \mathbb{C}^n \times \dots \times (\mathbb{C}^n)^*$, define $g_{(v, \phi)}: \mathrm{GL}(n, \mathbb{C}) \rightarrow \mathbb{C}$ by

$$g_{(v, \phi)}(A) = f(A \cdot (v, \phi)) - f((v, \phi)).$$

Then g is complex analytic, and g vanishes on $\mathrm{SU}(n)$ by hypothesis. So by Lemma 62, g vanishes on $\mathrm{SL}(n, \mathbb{C})$. Since this is true for each (v, ϕ) , it follows that f is invariant under $\mathrm{SL}(n, \mathbb{C})$. ♦

Now we really want to consider \mathbb{R} -valued functions of m vectors in \mathbb{C}^n which are polynomials in the real and imaginary parts of the components of the vectors. Actually, we might as well consider complex-valued functions which are polynomials over \mathbb{C} in the real and imaginary parts of the components of the vectors. Equivalently, we consider functions

$$f: \mathbb{C}^n \times \dots \times \mathbb{C}^n \rightarrow \mathbb{C}$$

which are polynomials over \mathbb{C} in the components v_{ri} of the vectors and in their complex conjugates \bar{v}_{ri} . Given such a function f , we define a function

$$\tilde{f}: \underbrace{\mathbb{C}^n \times \dots \times \mathbb{C}^n}_m \times \underbrace{(\mathbb{C}^n)^* \times \dots \times (\mathbb{C}^n)^*}_m \rightarrow \mathbb{C}$$

as follows:

- (i) if $f(v_1, \dots, v_m) = v_{ri}$, then $\tilde{f}(v_1, \dots, v_m, \xi_1, \dots, \xi_m) = v_{ri}$
- (ii) if $f(v_1, \dots, v_m) = \bar{v}_{ri}$, then $\tilde{f}(v_1, \dots, v_m, \xi_1, \dots, \xi_m) = \xi_{ri}$
- (iii) the correspondence $f \mapsto \tilde{f}$ is an algebra homomorphism.

* Actually, the exponential map is onto $\mathrm{SL}(n, \mathbb{C})$, but we won't prove that here.

Notice that \tilde{f} is a polynomial function in the (complex) components of the vectors of \mathbb{C}^n and $(\mathbb{C}^n)^*$. The mapping $f \mapsto \tilde{f}$ is clearly a one-one correspondence between the polynomials in the v_{ri} and \bar{v}_{ri} , and the polynomials in the v_{ri} and ξ_{ri} . Note that if f is the Hermitian inner product ι_{rs} of v_r and v_s , then \tilde{f} is the evaluation ε_{rs} . If f is the determinant of v_{r_1}, \dots, v_{r_n} , then \tilde{f} is also this determinant; if f is the conjugate of this determinant, then \tilde{f} is the determinant of $\xi_{r_1}, \dots, \xi_{r_n}$.

Suppose that $f(v_1, \dots, v_m) = \bar{v}_{ri}$. Then

$$\begin{aligned} (1) \quad f(A \cdot (v_1, \dots, v_m)) &= \text{the conjugate of the } i^{\text{th}} \text{ component of } v_r \cdot A^t \\ &= \text{the conjugate of } \sum_j a_{ij} v_{rj} \\ &= \sum_j \bar{a}_{ij} \bar{v}_{rj}. \end{aligned}$$

On the other hand, if $A \in U(n)$, so that $A^{-1} = \bar{A}^t$, then

$$\begin{aligned} (2) \quad \tilde{f}(A \cdot (v_1, \dots, v_m, \xi_1, \dots, \xi_m)) \\ &= \tilde{f}(v_1 \cdot A^t, \dots, v_m \cdot A^t, (A^{-1})^t \cdot \xi_1, \dots, (A^{-1})^t \cdot \xi_m) \\ &= i^{\text{th}} \text{ component of } (A^{-1})^t \cdot \xi_r \\ &= i^{\text{th}} \text{ component of } \bar{A} \cdot \xi_r \\ &= \sum_j \bar{a}_{ij} \xi_{rj}. \end{aligned}$$

Comparing (1) and (2), we see that

$$(3) \quad \tilde{f} \circ (A \cdot) = \widetilde{f \circ (A \cdot)},$$

where $(A \cdot)$ on the left is the action of $A \in U(n)$ on $\mathbb{C}^n \times \dots \times (\mathbb{C}^n)^*$, while $(A \cdot)$ on the right is the action of A on $\mathbb{C}^n \times \dots \times \mathbb{C}^n$. From the way that $f \mapsto \tilde{f}$ is defined, it is clear that (3) holds for all f . Consequently,

f is invariant under $U(n)$ if and only if \tilde{f} is invariant under $U(n)$.

From this we immediately conclude

64. THEOREM. Let $f: \mathbb{C}^n \times \dots \times \mathbb{C}^n \rightarrow \mathbb{C}$ be a polynomial function in the components of the vectors of \mathbb{C}^n and the conjugates of the components which is invariant under $SU(n)$. Then f is a polynomial in the Hermitian inner products and the determinants \det_{r_1, \dots, r_n} and their conjugates $\overline{\det_{r_1, \dots, r_n}}$.

PROOF. By Corollary 63, the function \tilde{f} is a polynomial in the evaluations ε_{rs} and the determinants \det_{r_1, \dots, r_n} and \det_{r_1, \dots, r_n}^* . Since

$$\varepsilon_{rs} = \widetilde{t_{rs}}, \quad \det_{r_1, \dots, r_n} = \overline{\det_{r_1, \dots, r_n}}, \quad \det_{r_1, \dots, r_n}^* = \overline{\det_{r_1, \dots, r_n}},$$

the result follows. ♦

In order to prove Theorem 49, which gives the corresponding result for $U(n)$, we need just one more observation. Let $v_1, \dots, v_n, w_1, \dots, w_n \in \mathbb{C}^n$, and let $A = (v_{ij}), B = (w_{ij})$. Then we have

$$(A \cdot \bar{B}^t)_{ij} = \sum_k v_{ik} \overline{w_{jk}} = \langle v_i, w_j \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the Hermitian inner product. Hence

$$(*) \quad \det A \overline{\det B} = \det(\langle v_i, w_j \rangle).$$

PROOF OF THEOREM 49. Since f is invariant under $SU(n)$, by Theorem 64 it can be written as a polynomial in the Hermitian inner products and

$$\det_{r_1, \dots, r_n} \quad \text{and} \quad \overline{\det_{s_1, \dots, s_n}}.$$

The action of an element $A \in U(n)$ multiplies the latter two by

$$\det A \quad \text{and} \quad \overline{(\det A)} = \det \bar{A} = \det(A^{-1})^t = (\det A)^{-1}.$$

Hence every factor \det_{r_1, \dots, r_n} must come paired with a factor $\overline{\det_{s_1, \dots, s_n}}$. But (*) says that the product of these two functions can be expressed in terms of the Hermitian inner product functions. ♦

ADDENDUM 2

RECOVERING THE DIFFERENTIAL FORMS;
THE GAUSS-BONNET-CHERN THEOREM
FOR MANIFOLDS-WITH-BOUNDARY

The crucial step in our proof of the generalized Gauss-Bonnet theorem was Theorem 22, for it immediately allowed us to conclude that if $\xi = \pi: E \rightarrow M$ is an oriented n -dimensional vector bundle, with sphere bundle $\pi_0: S \rightarrow M$, then $\pi_0^*C(\xi) = 0$. This means that if Λ is the n -form on M representing $C(\xi)$, then the n -form $\pi_0^*\Lambda$ on S is exact,

$$\pi_0^*\Lambda = d\Phi \quad \text{for some } (n-1)\text{-form } \Phi \text{ on } S.$$

In particular, suppose that $\xi = TM^n$, and let X be a unit vector field on M with a single isolated singularity, at $p \in M$ (Problem I.11-13). Let $B(\varepsilon)$ be a closed ball of radius ε around p , and set

$$M_\varepsilon = M - \text{interior } B(\varepsilon).$$

Then $X(M_\varepsilon) \subset S$ is a manifold-with-boundary, the image of M_ε under the section $X: M - \{p\} \rightarrow S$. Now

$$\begin{aligned} \int_M \Lambda &= \int_{M-\{p\}} \Lambda = \lim_{\varepsilon \rightarrow 0} \int_{M_\varepsilon} \Lambda = \lim_{\varepsilon \rightarrow 0} \int_{M_\varepsilon} X^*(\pi_0^*\Lambda) \\ &= \lim_{\varepsilon \rightarrow 0} \int_{X(M_\varepsilon)} \pi_0^*\Lambda = \lim_{\varepsilon \rightarrow 0} \int_{X(M_\varepsilon)} d\Phi \\ &= \lim_{\varepsilon \rightarrow 0} \int_{\partial X(M_\varepsilon)} \Phi. \end{aligned}$$

Recalling the definition of the index of X at p , we easily see that

$$\begin{aligned} (1) \quad \int_M \Lambda &= (\text{index of } X \text{ at } p) \cdot \int_{\pi_0^{-1}(p)} \Phi \\ &= \chi(M) \cdot \int_{\pi_0^{-1}(p)} \Phi \quad \text{by Theorem I.11-30.} \end{aligned}$$

Since for $n = 2m$ we also have

$$(2) \quad \int_M \Lambda = \int_M n! K_n dV = \pi^m m! 2^n \chi(M) \quad \text{by Theorem 26,}$$

we obtain, finally,

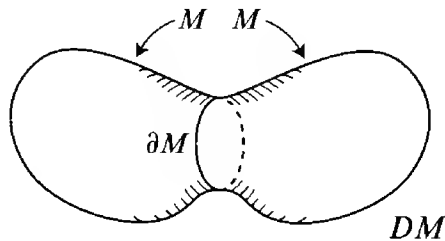
$$(3) \quad \int_{\pi_0^{-1}(p)} \Phi = \pi^m m! 2^n.$$

In the original intrinsic proof of the generalized Gauss-Bonnet theorem, Chern [2] did not use Theorem 22 or Corollaries 23 and 24. Instead, by clever guess-work he explicitly constructed a form Φ with $\pi_0^* \Lambda = d\Phi$, and noted that it satisfied equation (3). By applying equation (1), he thus deduced equation (2), which is precisely the generalized Gauss-Bonnet theorem. As we will soon see, it is very useful to have an explicit formula for Φ when we seek a generalized Gauss-Bonnet theorem for manifolds-with-boundary.

Let $(M, \partial M)$ be a compact orientable manifold-with-boundary. The Euler characteristic $\chi(M)$ is defined, as before, by

$$\chi(M) = \dim H^0(M) - \dim H^1(M) + \dots.$$

With some work, we could generalize Theorem I.11-5, and show that $\chi(M) = \alpha_0 - \alpha_1 + \dots$, where α_k is the number of k -simplexes in a triangulation. But we won't pause to prove this, because other facts about $\chi(M)$ are more important for us. First note that we can construct a compact oriented manifold DM , the **double** of M , by taking two disjoint copies of M , and identifying the corresponding points of ∂M . The following result is obvious in terms of



triangulations, but we will give an independent proof.

65. PROPOSITION. The Euler characteristic of DM is given by

$$\chi(DM) = 2\chi(M) - \chi(\partial M).$$

PROOF. Let U and V be open neighborhoods of the two copies of M in DM such that $H^k(U) \approx H^k(V) \approx H^k(M)$ for all k , and $H^k(U \cap V) \approx H^k(\partial M)$ for all k . Then we have the Mayer-Vietoris sequence (Theorem I.11-3)

$$\begin{aligned} 0 \rightarrow H^0(DM) \rightarrow \dots \rightarrow H^k(DM) \rightarrow H^k(U) \oplus H^k(V) \rightarrow \\ \rightarrow H^k(U \cap V) \xrightarrow{\delta} H^{k+1}(DM) \rightarrow \dots \end{aligned}$$

When we apply Proposition I.11-4, we obtain precisely the desired result. ♦

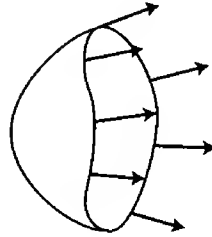
This result says quite different things when the dimension n of M is odd or even. For odd n we have $\chi(DM) = 0$ (Corollary I.11-25), so we obtain

$$\chi(M) = \frac{1}{2}\chi(\partial M);$$

in particular $\chi(\partial M)$ must be even. For even n , we have $\chi(\partial M) = 0$, so we have

$$(*) \quad 2\chi(M) = \chi(DM).$$

66. COROLLARY. Let M be a compact orientable manifold-with-boundary, of even dimension n . Let X be a vector field on M with only finitely many zeros, all in $M - \partial M$, such that X is outward pointing on ∂M . Then the sum of the indices of X is $\chi(M)$.



PROOF. We can modify X near ∂M so that X is the outward pointing unit normal ν on ∂M (and so that there are no new zeros). Then there is a vector field on DM which looks like X on one copy of M and like $-X$ on the other copy. Since n is even, the index of $-X$ at an isolated zero is the same as the index of X at that zero (Problem I.11-12). Thus Theorem I.11-30 gives

$$2(\text{sum of the indices of } X) = \chi(DM) = 2\chi(M), \quad \text{by } (*). \quad \spadesuit$$

67. COROLLARY. Let M be a compact oriented Riemannian manifold-with-boundary, of even dimension $n = 2m$, with tangent bundle $\pi: TM \rightarrow M$, and associated sphere bundle $\pi_0 = \pi|_S: S \rightarrow M$. Let ω be a connection on the principal bundle $\varpi: \text{SO}(TM) \rightarrow M$, with curvature form Ω , let Λ be the unique n -form on M with

$$\varpi^* \Lambda = \sum \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}} = 2^m m! \text{Pf}(\Omega),$$

and let Φ be an $(n-1)$ -form on S with

$$\pi_0^* \Lambda = d\Phi.$$

Finally, let $\nu: \partial M \rightarrow S$ be the outward pointing unit normal on ∂M . Then

$$\int_M K_n dV = \frac{1}{n!} \int_M \Lambda = \frac{\pi^m m! 2^n}{n!} \chi(M) + \frac{1}{n!} \int_{\partial M} \nu^* \Phi.$$

PROOF. Extend ν to a vector field X on M with only finitely many zeros $p_1, \dots, p_k \in M - \partial M$. Let $B_i(\varepsilon)$ be closed balls of radius ε around p_i which are disjoint from each other and from ∂M , and set

$$M_\varepsilon = M - \bigcup_{i=1}^k \text{interior } B_i(\varepsilon).$$

Then, as on page 380, we have

$$\begin{aligned} \int_M \Lambda &= \lim_{\varepsilon \rightarrow 0} \int_{\partial X(M_\varepsilon)} \Phi \\ &= \int_{\nu(\partial M)} \Phi + \sum_{i=1}^k \lim_{\varepsilon \rightarrow 0} \int_{\partial B_i(\varepsilon)} \Phi \\ &= \int_{\partial M} \nu^* \Phi + \sum_{i=1}^k \pi^m m! 2^n \cdot (\text{index of } X \text{ at } p_i) \quad \text{by (3), on page 381} \\ &= \int_{\partial M} \nu^* \Phi + \pi^m m! 2^n \cdot \chi(M) \quad \text{by Corollary 66. } \spadesuit \end{aligned}$$

The only trouble with this result is that we can't interpret $\int_{\partial M} \nu^* \Phi$ until we have an explicit Φ with $\pi_0^* \Lambda = d\Phi$. Fortunately, we can pull a Φ out of the air by looking more carefully at the proofs in section 4.

First consider two connections on the principal bundle $\text{SO}(E)$, as in Proposition 20. Changing notation slightly, we denote these connections by ω and $\tilde{\omega}$, with curvature forms Ω and $\tilde{\Omega}$. Let Λ and $\tilde{\Lambda}$ be the forms with

$$\varpi^*(\Lambda) = 2^m m! \text{Pf}(\Omega) \quad \text{and} \quad \varpi^*(\tilde{\Lambda}) = 2^m m! \text{Pf}(\tilde{\Omega}).$$

All quantities associated with $M \times [0, 1]$ will be written boldface, so the induced connections $q^* \omega$ and $q^* \tilde{\omega}$ in the proof of Proposition 20 will be denoted by

$$\omega = q^* \omega \quad \text{and} \quad \tilde{\omega} = q^* \tilde{\omega},$$

and we will set

$$\Psi = (1 - \tau)\omega + \tau\tilde{\omega}.$$

Note that $\partial/\partial\tau$ is horizontal for these connections.

The proof of Proposition 20 tells us how to find a form Φ with $d\Phi = \tilde{\Lambda} - \Lambda$. To obtain Φ explicitly, we first want to describe the curvature form Ψ of ψ more explicitly. Note that tangent vectors on $M \times [0, 1]$ can be considered as sums

$$X + \mu \frac{\partial}{\partial \tau} \quad \mu \in \mathbb{R}$$

where X is a tangent vector of M , and tangent vectors on the total space of $q^* \text{SO}(\xi)$ can be considered as sums

$$Y + \mu \frac{\partial}{\partial \tau}$$

where Y is a tangent vector on $\text{SO}(E)$. We have

$$d\psi = (1 - \tau)d\omega + \tau d\tilde{\omega} + d\tau \wedge (\tilde{\omega} - \omega),$$

so for two tangent vectors $Y_1 + \mu_1 \partial/\partial \tau$ and $Y_2 + \mu_2 \partial/\partial \tau$ on the total space of $q^* \text{SO}(\xi)$ we have

$$\begin{aligned} \Psi(Y_1 + \mu_1 \partial/\partial \tau, Y_2 + \mu_2 \partial/\partial \tau) &= d\psi(h(Y_1 + \mu_1 \partial/\partial \tau), h(Y_2 + \mu_2 \partial/\partial \tau)) \\ &= [(1 - \tau)d\omega + \tau d\tilde{\omega}](h(Y_1) + \mu_1 \partial/\partial \tau, h(Y_2) + \mu_2 \partial/\partial \tau) \\ &\quad + [d\tau \wedge (\tilde{\omega} - \omega)](Y_1 + \mu_1 \partial/\partial \tau, Y_2 + \mu_2 \partial/\partial \tau), \\ &\text{since } \tilde{\omega} - \omega = 0 \text{ on vertical vectors.} \end{aligned}$$

Extending Y_2 to a vector field, we have

$$d\omega(\partial/\partial \tau, h(Y_2)) = \frac{\partial}{\partial \tau}(\omega(h(Y_2)) - h(Y_2)(\omega(\partial/\partial \tau)) - \omega([\partial/\partial \tau, h(Y_2)]).$$

The first term on the right must vanish, since we can choose the vector field Y_2 to be independent of τ ; the second vanishes, since $\partial/\partial \tau$ is horizontal for ω ; and it is easily seen that the third term also vanishes (compare pg. IV.370). So if Ω and $\tilde{\Omega}$ are the curvature forms for ω and $\tilde{\omega}$, then we have, finally,

$$\begin{aligned} \Psi &= (1 - \tau)\Omega + \tau\tilde{\Omega} + d\tau \wedge (\tilde{\omega} - \omega) \\ &= \Omega + \tau(\tilde{\Omega} - \Omega) + d\tau \wedge (\tilde{\omega} - \omega). \end{aligned}$$

From this expression we see that

$$\begin{aligned} 2^m m! \text{Pf}(\Psi) &= \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Psi_{i_2}^{i_1} \wedge \dots \wedge \Psi_{i_n}^{i_{n-1}} \\ &= \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_n}^{i_{n-1}} \\ &\quad + d\tau \wedge \mathbf{H}, \end{aligned}$$

where \mathbf{H} is a linear combination of terms of the form

$$(*) \quad \tau^{m-k-1} \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_{2k}}^{i_{2k-1}} \wedge (\tilde{\Omega} - \Omega)_{i_{2k+2}}^{i_{2k+1}} \wedge \dots \wedge \\ \wedge (\tilde{\Omega} - \Omega)_{i_{n-2}}^{i_{n-3}} \wedge (\tilde{\omega} - \omega)_{i_n}^{i_{n-1}}.$$

We won't bother keeping track of the exact coefficients involved in our calculations, since there will be a cheap way of getting them out at the end. Our expression for $2^m m! \text{Pf}(\Psi)$ shows that the closed n -form Λ on $M \times [0, 1]$ which pulls back to $2^m m! \text{Pf}(\Psi)$ can be written

$$\Lambda = \dots + d\tau \wedge \eta,$$

where η is a linear combination of forms which pull back* to the forms $(*)$. Now Theorem I.7-17 says that

$$\tilde{\Lambda} - \Lambda = i_1^* \Lambda - i_0^* \Lambda = d(I\Lambda),$$

where

$$I\Lambda(p)(X_1, \dots, X_{n-1}) = \int_0^1 \eta(p, t)(i_{t*} X_1, \dots, i_{t*} X_{n-1}) dt.$$

In this integral, t will enter only in the factors t^{m-k-1} . All other terms are independent of t , since the connections ω and $\tilde{\omega}$ are independent of t —for a tangent vector Y on $\text{SO}(E)$ we have $\omega(i_{t*} Y) = \omega(Y)$, and similarly for $\tilde{\omega}$, Ω , and $\tilde{\Omega}$. So we see, finally, that

(A) $\tilde{\Lambda} - \Lambda = d\Phi$, where the $(n-1)$ -form Φ on M is a linear combination of $(n-1)$ -forms which pull back to the forms

$$(**) \quad \sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_{2k}}^{i_{2k-1}} \wedge (\tilde{\Omega} - \Omega)_{i_{2k+2}}^{i_{2k+1}} \wedge \dots \wedge \\ \wedge (\tilde{\Omega} - \Omega)_{i_{n-2}}^{i_{n-3}} \wedge (\tilde{\omega} - \omega)_{i_n}^{i_{n-1}}$$

on $\text{SO}(E)$.

It is easy to see that there are forms with this property, since Ω , $\tilde{\Omega}$, and $\tilde{\omega} - \omega$ vanish on vertical vectors. The proof is similar to that of Proposition 18, except that in the second part we explicitly write out the value of the form $()$ on $R_A^* Y_1, \dots, R_A^* Y_{n-1}$, noting that $R_A^* \omega = A^{-1} \omega A$ and $R_A^* \tilde{\omega} = A^{-1} \tilde{\omega} A$, by the definition of a connection. Then we check that this is $\det A = 1$ times the value of the form on Y_1, \dots, Y_{n-1} , the computation being similar to that in the proof of Proposition 9.

Now we will apply this to a special, complicated, case. Let $\xi = \pi: E \rightarrow M$ be an n -dimensional vector bundle with a Riemannian metric $\langle \cdot, \cdot \rangle$, let $\text{SO}(\xi) = \varpi: \text{SO}(E) \rightarrow M$ be the corresponding principal bundle, and let $\pi_0: S \rightarrow M$ be the corresponding sphere bundle. Set

$$\zeta = \pi_0^* \xi = p: \mathbf{E} \rightarrow M;$$

the total space \mathbf{E} consists of all pairs (e, v) where $e \in S$ and $v \in \pi(e)$. For the corresponding principal bundle

$$\text{SO}(\zeta) = \rho: \text{SO}(\mathbf{E}) \rightarrow S,$$

the fibre over e is the set of all (e, u) where $u \in \varpi^{-1}(\pi(e))$ is an orthonormal frame at $\pi(e)$. The principal bundle map

$$\begin{array}{ccc} \text{SO}(\mathbf{E}) & \xrightarrow{\tilde{\pi}_0} & \text{SO}(E) \\ \downarrow \rho & & \downarrow \varpi \\ S & \xrightarrow{\pi_0} & M \end{array}$$

$\tilde{\pi}_0: \text{SO}(\mathbf{E}) \rightarrow \text{SO}(E)$

which covers π_0 takes (e, u) to u .

Recall that we can write

$$\zeta = \zeta_1 \oplus \zeta_2,$$

where

$\zeta_1 = p_1: \mathbf{E}_1 \rightarrow M$ is an $(n-1)$ -dimensional bundle,

$\zeta_2 = p_2: \mathbf{E}_2 \rightarrow M$ is a trivial 1-dimensional bundle;

the total space $\mathbf{E}_2 \subset \mathbf{E}$ consists of all pairs $(e, \mu e)$ for $e \in S$ and $\mu \in \mathbb{R}$, while $\mathbf{E}_1 \subset \mathbf{E}$ consists of all pairs (e, v) where $v \in \pi(e)$ is orthogonal to e . For the corresponding principal bundles

$$\text{SO}(\zeta_i) = \rho_i: \text{SO}(\mathbf{E}_i) \rightarrow S,$$

the fibre $\rho_2^{-1}(e)$ contains just the two pairs (e, e) and $(e, -e)$, while the fibre $\rho_1^{-1}(e)$ consists of all pairs (e, v) where v is an ordered $(n-1)$ -tuple of orthonormal vectors at $\pi(e)$ all of which are perpendicular to e . We have a natural inclusion

$$\iota: \text{SO}(\mathbf{E}_1) \rightarrow \text{SO}(\mathbf{E}),$$

which takes (e, v) to (e, u) where u is the frame whose first $n-1$ members come from v , while $u_n = e$. Clearly,

$$\begin{array}{ccc} \text{SO}(\mathbf{E}_1) & \xrightarrow{\iota} & \text{SO}(\mathbf{E}) \\ & \searrow p_1 & \downarrow p \\ & & S \end{array}$$

$p \circ \iota = p_1.$

Now let ω be a connection on $\text{SO}(E)$, with curvature form Ω . Then

$$\gamma = \tilde{\pi}_0^* \omega$$

is a connection on $\text{SO}(\mathbf{E})$, whose curvature form Γ is $\tilde{\pi}_0^* \Omega$. Hence if Λ is the unique form on M with

$$\varpi^* \Lambda = 2^m m! \text{Pf}(\Omega),$$

then

$$(1) \quad p^* \pi_0^* \Lambda = \tilde{\pi}_0^* \varpi^* \Lambda = 2^m m! \text{Pf}(\tilde{\pi}_0^* \Omega) = 2^m m! \text{Pf}(\Gamma).$$

We can also use γ to determine a connection γ_1 on $\text{SO}(\mathbf{E}_1)$ by

$$(\gamma_1)_j^i = \iota^* \gamma_j^i \quad i, j = 1, \dots, n-1.$$

The two conditions

$$\begin{aligned} \gamma_1(\sigma(M)) &= M & \text{for } M \in \mathfrak{o}(n-1) \\ R_A^* \gamma_1 &= \text{Ad}(A^{-1}) \gamma_1 & \text{for } A \in \text{SO}(n-1) \end{aligned}$$

follow from the corresponding conditions for γ —if we regard $\text{SO}(\mathbf{E}_1) \subset \text{SO}(\mathbf{E})$ via the map ι , then $\sigma(M)$ in $\text{SO}(\mathbf{E}_1)$ is the restriction of $\sigma(M)$ in $\text{SO}(\mathbf{E})$ for $M \in \mathfrak{o}(n-1)$, and R_A in $\text{SO}(\mathbf{E}_1)$ is the restriction of $R_{\tilde{A}}$ in $\text{SO}(\mathbf{E})$, where $\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}$.

Although $(\gamma_1)_j^i = \iota^* \gamma_j^i$ for $i, j = 1, \dots, n-1$, it does not follow for these same values of i and j that the curvature forms Γ_1 and Γ satisfy $(\Gamma_1)_j^i = \iota^* \Gamma_j^i$; for the horizontal component in $\text{SO}(\mathbf{E}_1)$ is different from the horizontal component in $\text{SO}(\mathbf{E})$. To find the correct relationship, we use the second structural equation (Theorem II.8-16),

$$d\gamma(Y_1, Y_2) = -[\gamma(Y_1), \gamma(Y_2)] + \Gamma(Y_1, Y_2)$$

for tangent vectors Y_1, Y_2 on $\text{SO}(\mathbf{E})$, which means that

$$(2) \quad d\gamma_j^i(Y_1, Y_2) = - \sum_{k=1}^n \gamma_k^i(Y_1) \gamma_j^k(Y_2) - \gamma_k^i(Y_2) \gamma_j^k(Y_1) + \Gamma_j^i(Y_1, Y_2).$$

Similarly, for tangent vectors Z_1, Z_2 on $\text{SO}(\mathbf{E}_1)$, we have

$$d(\gamma_1)_j^i(Z_1, Z_2) = - \sum_{k=1}^{n-1} (\gamma_1)_k^i(Z_1) (\gamma_1)_j^k(Z_2) - (\gamma_1)_k^i(Z_2) (\gamma_1)_j^k(Z_1) + (\Gamma_1)_j^i(Z_1, Z_2)$$

which implies

$$(3) \quad d\gamma_j^i(\iota_* Z_1, \iota_* Z_2) = - \sum_{k=1}^{n-1} \gamma_k^i(\iota_* Z_1) \gamma_j^k(\iota_* Z_2) - \gamma_k^i(\iota_* Z_2) \gamma_j^k(\iota_* Z_1) + (\Gamma_1)_j^i(Z_1, Z_2).$$

Comparing (2) and (3), we see that

$$(4) \quad (\Gamma_1)_j^i = \iota^* \Gamma_j^i + \iota^*(\gamma_n^i \wedge \gamma_n^j) \quad i, j = 1, \dots, n-1.$$

Now we are going to apply the construction in the proof of Theorem 22. The principal bundle $\text{SO}(\mathbf{E}_2)$ for the 1-dimensional bundle ζ_2 is just 2 copies of M ; the only connection on $\text{SO}(\mathbf{E}_2)$ is $\gamma_2 = 0$. The bundle $\text{SO}(\mathbf{E}_1) * \text{SO}(\mathbf{E}_2) \subset \text{SO}(\mathbf{E})$ in the proof of Theorem 22 is just 2 copies of $\text{SO}(\mathbf{E}_1)$; the first copy may be identified with $\iota(\text{SO}(\mathbf{E}_1))$. The connection

$$\bar{\gamma} = \rho_1^* \gamma_1 \oplus \rho_2^* \gamma_2$$

on $\text{SO}(\mathbf{E}_1) * \text{SO}(\mathbf{E}_2)$ which is constructed in the proof of Theorem 22 is just $\iota^* \gamma_1$ on the first copy, and similarly the curvature form $\bar{\Gamma}$ is just $\iota^* \Gamma_1$ on the first copy. As before, we extend the connection $\bar{\gamma}$ to a connection $\tilde{\gamma}$ on $\text{SO}(\mathbf{E})$, with curvature form $\tilde{\Gamma}$. We have unique forms $\Upsilon, \tilde{\Upsilon}$ on S with

$$\begin{aligned} p^* \Upsilon &= 2^m m! \text{Pf}(\Gamma) \\ p^* \tilde{\Upsilon} &= 2^m m! \text{Pf}(\tilde{\Gamma}). \end{aligned}$$

But equation (1) says that $\Upsilon = \pi_0^* \Lambda$, while, as in the proof of Theorem 22, we have $\text{Pf}(\tilde{\Gamma}) = 0$ at points of $\text{SO}(\mathbf{E}_1) * \text{SO}(\mathbf{E}_2)$, which implies that $\tilde{\Upsilon} = 0$. Assertion (A) on page 385 thus shows that

$\pi_0^* \Lambda = d\Phi$, where the $(n-1)$ -form Φ on S is a linear combination of $(n-1)$ -forms which pull back, via p^* , to the forms

$$\sum_{i_1, \dots, i_n} \varepsilon^{i_1 \dots i_n} \Gamma_{i_2}^{i_1} \wedge \dots \wedge \Gamma_{i_{2k}}^{i_{2k-1}} \wedge (\tilde{\Gamma} - \Gamma)_{i_{2k}}^{i_{2k+1}} \wedge \dots \wedge (\tilde{\Gamma} - \Gamma)_{i_{n-2}}^{i_{n-3}} \wedge (\tilde{\gamma} - \gamma)_{i_n}^{i_{n-1}}.$$

But on tangent vectors to $\text{SO}(\mathbf{E}_1) * \text{SO}(\mathbf{E}_2) = 2$ copies of $\text{SO}(\mathbf{E}_1)$, we clearly have

$$(\tilde{\gamma} - \gamma)_j^i = \begin{cases} 0 & i, j \neq n \\ \iota^* \gamma_j^i & i \text{ or } j = n \end{cases} \quad \text{on the first copy,}$$

while equation (4) shows that we also have

$$(\tilde{\Gamma} - \Gamma)_j^i = \begin{cases} \iota^*(\gamma_n^i \wedge \gamma_n^j) & i, j \neq n \\ -\iota^* \Gamma_j^i & i \text{ or } j = n \end{cases} \quad \text{on the first copy.}$$

Hence, remembering that $\gamma = \tilde{\pi}_0^* \omega$ and $\Gamma = \tilde{\pi}_0^* \Omega$, we can conclude that

$\pi_0^* \Lambda = d\Phi$, where the $(n-1)$ -form Φ on S is a linear combination of $(n-1)$ -forms Φ_k , $0 \leq k \leq m-1$, such that

$$\begin{aligned} p_1^* \Phi_k &= \iota^* p^* \Phi_k \\ &= \iota^* \tilde{\pi}_0^* \left(\sum_{i_1, \dots, i_{n-1}} \varepsilon^{i_1 \dots i_{n-1}} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_{2k}}^{i_{2k-1}} \right. \\ &\quad \left. \wedge \omega_n^{i_{2k+1}} \wedge \dots \wedge \omega_n^{i_{n-1}} \right); \end{aligned}$$

in this sum, the indices i_α run from 1 to $n-1$.

This can be put in a more useful form by introducing the “last vector” map $\ell: \text{SO}(E) \rightarrow S$ defined by

$$\ell(u) = u_n.$$

We clearly have

$$\ell \circ \tilde{\pi}_0 \circ \iota = p_1: \text{SO}(E_1) \rightarrow S.$$

It is easy to check that there are unique forms Θ_k on S such that

$$\ell^* \Theta_k = \sum_{i_1, \dots, i_{n-1}} \varepsilon^{i_1 \dots i_{n-1}} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_{2k}}^{i_{2k-1}} \wedge \omega_n^{i_{2k+1}} \wedge \dots \wedge \omega_n^{i_{n-1}};$$

we use the procedure given in the footnote on page 385, noting that in the

second part we only want to consider $A \in \text{SO}(n)$ with $\ell(u \cdot A) = \ell(u)$, so that $A = \begin{pmatrix} B & 0 \\ 0 & 1 \end{pmatrix}$ for $B \in \text{SO}(n-1)$. But now we have

$$\begin{aligned} p_1^* \Phi_k &= \iota^* \tilde{\pi}_0^* \ell^* \Theta_k = p_1^* \Theta_k \\ \implies \Phi_k &= \Theta_k. \end{aligned}$$

So we can state, finally,

(B) There are constants a_0, \dots, a_{m-1} such that

$$\pi_0^* \Lambda = \sum_{k=0}^{m-1} a_k \Phi_k,$$

where the Φ_k are the unique forms on S such that

$$\ell^* \Phi_k = \sum_{i_1, \dots, i_{n-1}} \varepsilon^{i_1 \dots i_{n-1}} \Omega_{i_2}^{i_1} \wedge \dots \wedge \Omega_{i_{2k}}^{i_{2k-1}} \wedge \omega_n^{i_{2k+1}} \wedge \dots \wedge \omega_n^{i_{n-1}}.$$

Note that the constants a_0, \dots, a_{m-1} do not depend on the bundle, or anything else; they are certain combinatorial terms in a formal calculation which is exactly the same for all bundles. If we apply Corollary 67, we obtain

$$(*) \quad \int_M K_n dV = \frac{\pi^m m! 2^n}{n!} \chi(M) + \frac{1}{n!} \sum_{k=0}^{m-1} a_k \int_{\partial M} \nu^* \Phi_k.$$

If we choose a positively oriented orthonormal moving frame $\mathbf{X} = (X_1, \dots, X_n)$ on ∂M with $X_n = \nu$, then

$$\nu = \ell \circ \mathbf{X}, \quad \mathbf{X}: \partial M \rightarrow \text{SO}(E),$$

so

$$\begin{aligned} \nu^* \Phi_k &= \mathbf{X}^* \ell^* \Phi_k \\ &= \sum_{i_1, \dots, i_{n-1}} \varepsilon^{i_1 \dots i_{n-1}} (\mathbf{X}^* \Omega_{i_2}^{i_1}) \wedge \dots \wedge (\mathbf{X}^* \Omega_{i_{2k}}^{i_{2k-1}}) \\ &\quad \wedge (\mathbf{X}^* \omega_n^{i_{2k+1}}) \wedge \dots \wedge (\mathbf{X}^* \omega_n^{i_{n-1}}). \end{aligned}$$

Recall that ω and Ω are forms on $\text{SO}(E)$; the terms $\mathbf{X}^* \omega_n^{i_{2k+1}}$ and $\mathbf{X}^* \Omega_{i_2}^{i_1}$ are just the corresponding connection and curvature forms for the moving frame \mathbf{X} .

This allows us to give an invariant definition of $\nu^*\Phi_k$ similar to the invariant definition of $\Lambda/n! = K_n dV$: if dV_{n-1} denotes the volume element on ∂M , then

$$\nu^*\Phi_k = K_k dV_{n-1},$$

where K_k can be written as a contraction of tensor products of the tensor \mathfrak{e} on ∂M (contravariant of order $n-1$), the curvature tensor \mathcal{R} for ∂M , and the tensor

$$(X, Y) \mapsto \langle \nabla_X \nu, Y \rangle,$$

which is just the second fundamental form of ∂M in M .

To calculate the constants a_0, \dots, a_{m-1} , we just have to apply equation (*) to products $M = D^{k+1} \times S^{n-k-1}$, with $\partial M = S^k \times S^{n-k-1}$; then the only non-zero boundary integral is the one involving $\nu^*\Phi_k$. The explicit calculations are left to the reader—after doing them, it should be fun to compare with Chern's paper [2].

BIBLIOGRAPHY

Part A of the bibliography describes the main topics of differential geometry not included in these volumes, as well as many subsidiary topics. Part B lists books, monographs, etc., referred to by curly brackets { }, while Part C lists individual journal articles, referred to by square brackets [].

The list of journal articles is rudimentary, containing only a few articles not specifically referred to in the text; no serious attempt has been made to provide historical background or to attribute theorems justly. A more than adequate supply of references can be obtained from the extensive bibliographies in some of the books listed in Part B, and of course there's always *Mathematical Reviews*.

A. OTHER TOPICS IN DIFFERENTIAL GEOMETRY

I. Major topics everyone should know something about.*

(a) *Complex manifolds*. This is the main topic which was hardly touched upon in these notes. Some people heartily dislike the subject, with its penetrating algebraic odor, but for others it has a seductive appeal. A differential geometer whose work often uses the simplifications obtained by considering the complex domain explained to me that the additional structure of complex manifolds makes them more interesting, just as two sexes are more interesting than one, but various aspects of this argument are open to debate. A basic treatment of complex manifolds is given in Kobayashi and Nomizu {1; v.2, ch.9}. See also Chern {1}, Weil {1}, Wu {1}, and Yano {2}. For more emphasis on the analysis aspects, see Griffiths {1}, Morrow and Kodaira {1}, and Wells {1}.

(b) *Homogeneous spaces*. In Chapter 13 we defined a “homogeneous space” to be a quotient space G/H . The terminology comes from the fact that these are precisely the manifolds M on which G acts transitively (see Warner {1; 120ff.} or Wolf {1; 11–13}); one should also be aware of the usage in the case of Riemannian manifolds (cf. Kobayashi and Nomizu {1; v.1, 176}). Once the identification with G/H is made, further study of these spaces becomes rather algebraic. See Kobayashi and Nomizu {1; v.2, ch.10}. It should also be mentioned that homogeneous spaces provide the natural setting for geometry according to the famous definition proposed by Felix Klein [1], of geometry as the theory of geometric invariants of a transitive transformation group. Even such subjects as Riemannian geometry have been brought within the compass of this definition, essentially by means of connections on certain principal bundles. There are some older references at the end of Chapter 2 of Veblen and Whitehead {1}, but the most extensive treatment is given by Sharpe {1}.

(c) *Symmetric spaces*. A very nice brief introduction to (Riemannian) symmetric spaces can be found in Milnor {2; §§20, 21}. Increasing detail, and algebra, can be found in Wolf {1}, Kobayashi and Nomizu {1; v.2, ch.11}, Boothby and Weiss {1}, Loos {1}, and the standard treatise Helgason {1}, with specialized material in Eberlein {1}.

(d) *Mappings*. Numerous types of maps between Riemannian manifolds are of importance—*isometries*, *similarities* (which multiply the metrics by a constant), *conformal maps* (which multiply the metrics by a function), *affine maps* (which take geodesics into geodesics), and *projective maps* (which take geodesics into

* As evidenced by the fact that they get a chapter apiece in Kobayashi and Nomizu {1}.

reparameterized geodesics); the latter two can be defined for arbitrary connections. We might also mention [essential] volume preserving maps, which preserve volumes of open subsets [up to a constant factor]. Certain classical relations between such maps, not mentioned here, may be found, for example, in Laugwitz {1; 147–161}. In preference to Theorem 13.4.6, one may consult Lemma 1 on pg. 242 of Kobayashi and Nomizu {1; v.1} (Laugwitz defines “irreducible” incorrectly); their proof of Lemma 2 on the same page is perhaps also somewhat preferable to Laugwitz’s proof of the corresponding Theorem 13.6.2. One other classical result may be found in Haack {1; 130–133} or Kreyszig {1; 267–269}. Naturally, the study of maps from S^2 to \mathbb{R}^2 has received special attention. Although cartography is an independent subject, the reader will probably find more than enough information about it in classical differential geometry books, for example Kreyszig {1} and Laugwitz {1}, or Scheffers {1; v.2, 36–53} and Strubecker {1; v.2, 170–201} for more examples. In addition to the maps themselves, one can study vector fields which represent “infinitesimal” versions of them. In particular, the infinitesimal versions of isometries are the “Killing vector fields”. For basic information on mappings and infinitesimal mappings see Kobayashi and Nomizu {1; v.1, ch.6}, Lichnerowicz {2}, and Yano {3}.

II. Other topics of substantial interest.

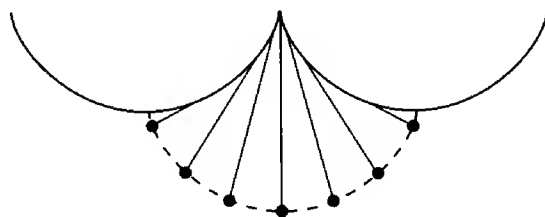
(a) *Classical curve and surface theory.* There is still a lot of information to be mined here, though the ore naturally tends to decrease in quality, rapidly passing the point of diminishing returns.

In our discussions of curve theory, we initially sought the limiting circle through 3 points on a plane curve. This osculating circle is often said to have second order contact with the curve. The notion of contact of curves and surfaces is described in Struik {1; 23}, somewhat more carefully in Goetz {1; 37, 44} and Kreyszig {1; 47–51, see especially Theorem 14.3}, and in detail in Favard {1; Part 1, ch.2}. Innocuous as the concept may seem, it is sometimes useful to have precise information about it (cf. do Carmo and Warner [1; pg. 136]).

In the theory of space curves it is natural to seek an “osculating sphere” having third order contact with the curve—see Blaschke {1; v.1, 33}, Eisenhart {2; 37}, Goetz {1; 77}, Kreyszig {1; 51}, Struik {1; 25}, or Gerretsen {1; 91}, which gives the analogous considerations in \mathbb{R}^n . The condition that a curve lie on some sphere is usually determined by setting equal the radii of all osculating spheres. Problem III.4-2 gives a different approach (generalized in Gerretsen {1; 78}); some calculation is required to establish the equivalence of the two answers.

A standard topic in curve theory is the study of involutes and evolutes, which was originated (cf. Coolidge {1; 319}) by Huygens in order to construct a pendu-

lum whose period is independent of its amplitude (for ordinary pendulums this is only approximately so, for small amplitudes). This property is possessed by a



pendulum whose weighted end describes a cycloid, hence the problem of finding a curve whose “involute”, traced out by unwinding a thread from it, will be a cycloid. The desired curve (the “evolute” of the cycloid) turns out to be another cycloid. [Unfortunately, this ingeniously designed Huygens pendulum was superseded by a pendulum suspended from a spring, which turns out to work just as well.] Two other familiar curves might be mentioned here: the evolute of the tractrix is a catenary. The standard material, none too interesting, on involutes and evolutes (which are also defined for space curves) can be found in Eisenhart {2; 43–45}, Gerretsen {1; 83–87}, Goetz {1; 65–70}, Kreyszig {1; 52–54}, Struik {1; 39–41}, and Strubecker {1; v.1, 222–226}. Guggenheimer {1; 35–47} gives a treatment requiring less differentiability, by means of the Riemann-Stieltjes integral (or see Ostrowski [1]), and finds all plane curves similar to their evolutes (pp. 59–61).

It is strange that the theory of envelopes is so seldom mentioned in connection with involutes and evolutes, for the evolute of a plane curve is the envelope of its normals, and thus the locus of the centers of its osculating circles. Even the latter fact is seldom mentioned (cf. Guggenheimer {1; 43–44}). As the figure on pg. II.6 seems to indicate, an osculating circle of a curve separates the parts of the curve with smaller curvature from those with larger curvature. A direct verification may be given (cf. Goetz {1; 84}), but it is even easier to prove a much stronger result, due to Kneser. If c is a curve, parameterized by arclength, with κ nowhere 0, then the curve of centers of the osculating circles is $\gamma(s) = c(s) + \mathbf{n}(s)/\kappa(s)$. The parameter s is not arclength for γ . Instead we have

$$\begin{aligned}\gamma'(s) &= c'(s) + \left(\frac{1}{\kappa}\right)'(s)\mathbf{n}(s) + \frac{1}{\kappa(s)}\mathbf{n}'(s) \\ &= c'(s) + \left(\frac{1}{\kappa}\right)'(s)\mathbf{n}(s) + \frac{1}{\kappa(s)} \cdot (-\kappa(s)c'(s)) \quad \text{by Serret-Frenet} \\ &= \left(\frac{1}{\kappa}\right)'(s)\mathbf{n}(s).\end{aligned}$$

Now κ is not constant on any interval, so if $\kappa' \neq 0$ everywhere, then γ' is not constant on any interval, and hence no portion of γ is a straight line. Therefore

$$|\gamma(s_1) - \gamma(s_2)| < \text{length of } \gamma \text{ from } \gamma(s_1) \text{ to } \gamma(s_2) \\ = \int_{s_1}^{s_2} |\gamma'(t)| dt = \left| \frac{1}{\kappa}(s_2) - \frac{1}{\kappa}(s_1) \right|,$$

where the left side is the distance between the centers of the osculating circles at s_1 and s_2 , while the right side is the difference in the radii of these osculating circles. Hence one must lie inside the other. Thus all the osculating circles are nested (and the smaller ones, containing points of the curve with larger curvature, must lie inside the larger ones). This gives a striking example of a family of curves (the osculating circles) none of which intersect, but which nevertheless have an envelope (the original curve).

There is a beautiful little book of Boltyanskii {1}, unfortunately out of print, which makes the study of envelopes seem very pretty. In differential geometry books, emphasis is usually placed on envelopes of families of surfaces; see Eisenhart {2; 59–65} and Kreyszig {1; 253–263}. There is a fairly thorough treatment of envelopes in Favard {1; Part 1, ch.3}. We should not fail to mention that the conjugate locus of a point is just the envelope of the geodesics through the point (refer again to the picture on pg. IV.221). An argument of Carathéodory shows that the conjugate locus always has at least 4 cusps (cf. Blaschke {1; v.1, 231}). Similar arguments show that the evolute (= envelope of the normals) of a closed plane curve must have at least 4 cusps. This actually follows immediately from the four vertex theorem, but it can also be used to prove the four vertex theorem, as well as a four vertex theorem for the hyperbolic plane (oral communication by A. Weinstein).

Naturally, many special sorts of space curves are investigated in the classical books. Special mention may be made of Bertrand curves—see Blaschke {1; v.1, 35}, Eisenhart {2; 39–41}, Gerretsen {1; 83}, Goetz {1; 74–76}, Haack {1; 29}, or Strubecker {1; v.1, 228–238}. They are of some interest, as they may be used to prove (Catalan's theorem) that the only ruled minimal surface is the helicoid—see do Carmo {1}. (This also follows from a classical result of Schwarz that a minimal surface containing a straight line is taken into itself by a rotation of π around the line. See, for example, Blaschke {1}.) Perhaps this is also a suitable place to mention an elementary theorem of Beltrami: the tangent developable of c intersects the osculating plane of c at $c(t)$ in a curve γ whose curvature at $c(t)$ is $3/4$ the curvature of c at this point.

Moving on to global theorems, we first mention that a simple proof of Theorem II.1-8 can be given when κ is nowhere 0 by noting that the curve is locally

convex, and then using a theorem of Schmidt (cf. Stoker {1; 46–47}) that local convexity implies convexity. (But for this proof we need to know that the curve bounds a region, to which Schmidt's theorem is applied. The proof given previously, using the “Hopf Umlaufsatz” (Theorem II.1-7), not only works for $\kappa \geq 0$ or $\kappa \leq 0$, but also proves directly that the curve lies on one side of each of its tangent lines, a criterion for convexity which does not use the fact that the curve bounds a region.) Schmidt's result holds in all dimensions, and could also be used to prove Hadamard's theorem (Theorem III.2-11) for imbedded surfaces.

Some theorems of Schwarz, Schur, and Schmidt are especially interesting because they are global theorems about non-closed curves—see Blaschke {1; v.1, 61–64}, as well as Chern {3; 35–38}. Guggenheimer {1; 31} proves one of these theorems in the special case where both curves are planar; Hilbert and Cohn-Vossen {1; 211–212} show how to obtain the general case from this, and then give some further discussion. Compare also Blaschke {1; v.1, §39}. Our old friend, the four vertex theorem, can be proved from the planar case (Guggenheimer {1; 30–32} or Fog [1]). By the way, in our proof of the four vertex theorem we only obtained 4 points where $\kappa' = 0$, but it is easy to actually obtain 4 relative maxima or minima of κ . Other interesting global theorems about non-closed curves are due to Vogt and Ostrowski (see Guggenheimer {1; 49–53} or Ostrowski [2]).

Laugwitz {1; 198–202} has some results on curves of constant width, which may also be proved by more elementary means—see, for example, the beautiful book of Yaglom and Boltyanskii {1; ch.7}.

Notice that the formula for τ on pg. III.225 can be written $\tau = (\arctan \kappa_g)'$. It follows from this that $\int \tau ds = 0$ for a closed curve lying on a sphere. Conversely, if this holds for all closed curves on a surface, then the surface is part of a plane or sphere (Scherrer [1]). The same results hold for $\int \kappa^n \tau ds$ (Saban [1]). Finally, we mention that any curve in S^2 is the unit tangent \mathbf{t} of some curve c of constant torsion. Such curves were studied classically (see, for example, Blaschke {1; v.1, 47}, and Darboux {1; §§36, 39 and v.4, Note 7, §7}), but only recently have closed curves of constant torsion been discovered (Weiner [1]; compare pg. IV.110).

Classical surface theory is of course much more extensive, and there is so much material contained in the standard classical books that there is no point trying to list the main topics. In Part B mention is sometimes made of specific information contained in particular books. For serious digging be sure not to forget the *Encyklopädie der Mathematischen Wissenschaften*.

See also the references under III.(c).

(b) *Extremal and isoperimetric problems.* Various extremal and isoperimetric problems for curves are treated in Blaschke {1; v.1, ch.2}. See Chapter 8 of the same volume for surfaces, and Chapters 2 and 6 of the second volume for similar problems in special affine geometry. For various sorts of solutions to the isoperimetric problem see also Blaschke and Reichardt {1; §28}, Chern {3; 25–29}, and Guggenheimer {1}. The last has a solution in the plane (pp. 79–84) which generalizes (pg. 289) to convex surfaces, by means of Steiner's formulas relating the area A and enclosed volume V of a compact convex surface $M \subset \mathbb{R}^3$ to the area $A(\varepsilon)$ and the enclosed volume $V(\varepsilon)$ of its parallel surface $\{p + \varepsilon \nu(p) : p \in M\}$:

$$A(\varepsilon) = A + 2\varepsilon \int_M H \, dA + 4\pi \varepsilon^2$$

$$V(\varepsilon) = V + \varepsilon A + \varepsilon^2 \int_M H \, dA + \frac{4}{3}\pi \varepsilon^3.$$

(The first formula follows from Problem III.3-12, and the second by integrating with respect to ε . It can also be proved by approximating the surface by convex polyhedra. In this case H measures dihedral angles and K measures vertex angles; for more details see pp. 168–170 of the article by Santalo in Chern {3}.) See also Blaschke {3} or Santalo {1} for a treatment of the isoperimetric problem by integral geometry (III.(a)). There is a detailed discussion of the isoperimetric problem in Blaschke {2}, but for the final word (including the isoperimetric problem in the spaces of constant curvature) see Hadwiger {1} and references therein. For later work, see Chavel {2}.

(c) *Closed geodesics.* For brief remarks on the existence of closed geodesics see Blaschke {1; v.1, 211–212}. See also pg. 233 for surfaces on which all geodesics are closed; for more details consult Berger {1}, which also proves the theorem of L. W. Green [1] that the sphere is the only surface on which every point has a unique conjugate point. For modern treatments of closed geodesics see Schwartz {1}, Flaschel and Klingenberg {1}, Besse {1}, and Klingenberg {1}, {2}.

(d) *Holonomy.* The holonomy group of a connection on a principal bundle P with group G is the subgroup of all $a \in G$ such that a fixed $u \in P$ can be joined to $u \cdot a$ by a horizontal curve. Thus, the holonomy group measures the extent to which the distribution of horizontal subspaces is not integrable. [In classical mechanics (V.(a)) a system of “constraints” which can be described by a suitable distribution is called “holonomic” if the distribution is integrable, and “non-holonomic” otherwise. Thus, the “holonomy group” really should be called the “non-holonomy group”, since it measures the extent to which a distribution is non-integrable.] Holonomy groups are studied in great detail in Kobayashi

and Nomizu {1}, with basic properties treated in chapter 2 of volume 1, and applications throughout. In particular, we have the “holonomy theorem” of Ambrose and Singer (first proved, or at least stated, by É. Cartan), which describes the Lie algebra of the holonomy group in terms of the curvature form of the connection. It should perhaps be pointed out that this is in some sense a global version of a classical description of the curvature tensor in terms of parallel translation around an “infinitesimal parallelogram”—see Eisenhart {1; 65}, Kreyszig {1; 295}, or Laugwitz {1; 108}, or slightly different versions in Bishop and Crittendon {1; 97}, Nelson {1; 77}, or Singer and Thorpe {1; 170–174}. It should also be noted that the holonomy theorem gives an immediate proof of the Test Case. The most important application of holonomy groups for Riemannian manifolds is the de Rham decomposition theorem (Kobayashi and Nomizu {1; v.1, 187 ff.}).

(e) *Reducing the group of a bundle; G -structures.* The proof of the holonomy theorem uses the concept of a reduction of the group G of a principal bundle P to a subgroup H . This is, by definition, a subset P' of P such that $u \cdot a \in P'$ for all $u \in P'$ and $a \in H$, so that P' is a principal bundle with group H . The prime example is a reduction of the group $GL(n, \mathbb{R})$ of the bundle of frames of M to the subgroup $O(n)$. Any Riemannian metric $\langle \cdot, \cdot \rangle$ gives such a reduction—we define P' to be the set of all frames which are orthonormal with respect to $\langle \cdot, \cdot \rangle$. Conversely, given any such reduction, we can define $\langle \cdot, \cdot \rangle$ by declaring the frames in P' to be orthonormal. Similarly, an orientation on M is equivalent to a reduction of the group $GL(n, \mathbb{R})$ to the group $GL^+(n, \mathbb{R})$ of matrices of positive determinant. It could hardly be supposed that mathematicians would not get around to generalizing these examples: a reduction of the bundle of frames on M to a subgroup $G \subset GL(n, \mathbb{R})$ is called a **G -structure**. For the theory of G -structures see the last chapter of Sternberg {1}, and Kobayashi {1}.

(f) *Contact transformations and contact structures.* At each point p of M we can consider the set of $(n - 1)$ -dimensional subspaces of the tangent space M_p . With the notation of Chapter V.13, this set would be denoted by $G_{n-1}(M_p)$; in the terminology of topic III.(g) it is the set of 1st order $(n - 1)$ -dimensional contact elements at p . We can form the manifold $C_{n-1}^1 M = \bigcup_{p \in M} G_{n-1}(M_p)$ of all these contact elements, and any immersion $f: M \rightarrow N$ gives rise to a map $f_*: C_{n-1}^1 M \rightarrow C_{n-1}^1 N$. An arbitrary smooth map $g: C_{n-1}^1 M \rightarrow C_{n-1}^1 N$ was classically called a **contact transformation** if it satisfies the following condition, which is automatic for f_* : for every hypersurface $P \subset M$, there is a hypersurface $Q \subset N$ such that the set of tangent spaces of Q is just the image under g of the set of tangent spaces of P . This geometric definition is unfortunately rather vague, since we want to allow the possibility, for example, that Q is a

single point $q \in N$ and g takes all tangent spaces of P into $G_{n-1}(N_q)$. But it is not hard to derive a precise analytic condition which captures the geometric content. The manifold $C_1^1\mathbb{R}^2$, for example, has a covering by two coordinate systems, one of which is defined on the set U of all directions not parallel to the y -axis—we use the coordinates a, b of the point $p \in \mathbb{R}^2$ as two coordinates on U and the slope m of the line in \mathbb{R}^2_p as the third coordinate. For a curve c in \mathbb{R}^2 we have

$$m(\text{subspace spanned by } c'(t)) = \frac{db(c'(t))}{da(c'(t))}.$$

From this it is not hard to see that a map $g: U \rightarrow U$ should be called a contact transformation if and only if $g^*(db - m da) = \alpha(db - m da)$ for a nowhere 0 function α . For $C_n^1\mathbb{R}^{n+1}$, with x^1, \dots, x^n, z as coordinates for the point, and y^i ($i = 1, \dots, n$) as the slope of the intersection of the n -dimensional subspace with the (x^i, z) -plane, we have the analogous criterion, in terms of the form $dz - \sum_i y^i dx^i$. For arbitrary manifolds these conditions can be formulated on coordinate neighborhoods. Although the classical reference Lie and Scheffers {2} will present problems, it is delightfully concrete and filled with examples; see also Eisenhart {4; ch.6} and Favard {1; part 1, ch.4}.

Nowadays, these motivating geometric considerations are almost never mentioned (an exception is Hermann {2; ch.3}). The modern approach to the subject may be found in Kobayashi and Nomizu {1; v.2, 381–382} and Kobayashi {1; 28 ff.}; it involves the notion of a **contact structure**, which also plays an important role in classical mechanics (V.(a)). An important tool in the study of contact structures is a theorem of Darboux, which is proved in Kobayashi {1; Appendix 1}; a very different proof is given in Lang {1; ch.5, §7}. The proof in Godbillion {1; 115–121} or Sternberg {1; 135–141} is of interest, as it uses the “characteristic system” of an ideal of differential forms. This “characteristic system” is related to the characteristics of a PDE; to see this made more explicit one may consult Dieudonné {1; v.4, 92–118}. The “Legendre transformation” is a contact transformation which is often used in PDE’s (see, for example, Courant and Hilbert {1; v.2, 32–29}). Legendre transformations are also used in the calculus of variations and classical mechanics (cf. Abraham {1}, Godbillion {1}, or Sternberg {1}). For the connection between the two, try Hermann {2; ch.6, §9}.

We could also consider maps $g: C_r^k M \rightarrow C_r^k N$, defined on k^{th} order r -dimensional contact elements [cf. III.(g)], satisfying an analogous geometric condition. But these are essentially of the form f_* for $r < n - 1$, or an extension of a contact transformation $g: C_{n-1}^1 M \rightarrow C_{n-1}^1 N$ for $r = n - 1$.

(Knebelman [1]). Similarly, one can define “infinitesimal contact transformations”, but they always come from a map $f: M \rightarrow N$. See Eisenhart {4; 252} or Lie and Scheffers {2; ch.4, §2} for a classical statement of this fact, and Kobayashi {1; 30} for a modern version.

(g) *The Laplacian and Hodge theory*. Berger, Gauduchon, and Mazet {1} is an excellent introduction to the significance of the Laplacian, though somewhat out of date because of the recent rapid progress in this field. For many applications of Hodge theory similar to Bochner’s Theorem (Theorem IV.7-63), see Yano and Bochner {1}, and Yano {1} (which also has applications of integral formulas similar to those in Chapter V.12). See also Ruse, Walker, and Wilmore {1}, and Jost {1}. A simple application to prove Poincaré duality is given in Warner {1}; for applications to complex manifolds see, for example, Weil {1}.

III. Other geometries

(a) *Finsler geometry*. For a brief introduction, see Laugwitz {1; §15}. For an extended treatment see Rund {1} and Matsumoto {1}. The reprinting of the thesis of Finsler {1}, with a bibliography up to that time, may be of interest.

(b) *Integral geometry*. Blaschke {3} and Santalo {1} are very nice introductions to this subject. The article by Santalo in Chern {3} gives references to later work. It is of interest to compare the arguments on pg. 167 with the proofs of Fenchel’s theorem and the Fary-Milnor theorem on pp. 33–35.

(c) *Line geometry*. Here one studies the manifold consisting of all lines in \mathbb{R}^3 . I haven’t the slightest idea what is done, but there are supposed to be some nice things. See Blaschke {1; v.1, ch.9}, Eisenhart {2; ch.12}, Favard {1; pt.2, sec.1, ch.5}, Forsyth {1; ch.12} and Hlavatý {2}. See also (e). The manifold of all circles in \mathbb{R}^3 has also been studied. See Eisenhart {2; ch.13} and Forsyth {1; ch.12}.

(d) *Affine geometry*. For special affine curve theory see Blaschke {1; v.2} and Favard {1; pt.2, sec.2, ch.1}, as well as Guggenheimer {1; §8-3}. For special affine surface theory see Blaschke {1; v.2} and Favard {1; pt.2, sec.2, ch.2}. See also Flanders [1] and the references listed under it. Not much seems to have been written about general affine invariants. See Guggenheimer {1; ch.7-3, probs.10–12} and Dieudonné {1; v.4, ch.20, §14, probs.11, 12}, which first describes an affine normal (“pseudo-normale”) for a hypersurface of any manifold with a torsion-free connection on its bundle of frames [compare with the second part of (e)] and then specializes to special affine geometry of \mathbb{R}^n (it will probably make things much easier to rewrite the problem so that it deals with moving frames, rather than with the bundle of frames itself). On the other hand, there

is a considerable literature on what might be called “general linear geometry”—properties of submanifolds of \mathbb{R}^n invariant under all linear maps (but not translations!), as well as properties invariant under all elements of $SL(n, \mathbb{R})$, again excluding translations [this certainly seems like a strange geometry to study, but see Laugwitz {2}]. See Salkowski {1}, Shirokow and Shirokow {1}, and Nomizu and Sasaki {1}.

(e) *Projective differential geometry*. This is the study of properties of submanifolds of projective space \mathbb{P}^n which are invariant under the projective group (cf. Harts-horne {1} for basic terminology). An eminent differential geometer, who perhaps prefers to remain anonymous, has said that the problem with projective differential geometry is that the projective group is too large to allow any interesting local results, while no one has ever discovered any interesting global ones.

The best introduction is probably Fubini and Čech {1}, which also introduces É. Cartan’s methods (using moving frames—compare (g)) as worked out in Cartan {5}. Other texts are Bol {1}, Favard {1; pt.2, sec.3}, Lane {1}, Akivis and Goldberg {1}, and Wilczynski {1}, one of the earliest works in the field [cf. (g)]. For “line geometry” in the projective case see Švec {1}.

Just as Riemannian geometry generalizes the differential geometry of \mathbb{R}^n , so one might expect to generalize projective differential geometry to an arbitrary manifold M by forming the union of the projective spaces obtained from each tangent space M_p , constructing a corresponding principal bundle P , and considering some canonical connection ω on P . All these steps can be carried out, but ω is not characterized so simply as in the Riemannian case by having vanishing torsion (which is defined only for connections on the bundle of frames); instead, there is a unique ω which satisfies certain identities like the Bianchi identities. This is described in Cartan {5}, but it will probably be much easier to read Kobayashi {1; 127–138}.

(f) *Other esoteric geometries*. See Blaschke {1; v.3} for the interesting, but complicated, geometries of Möbius, Laguerre, and Lie; for the last of these, see also Cecil {1}. In all three geometries, the points are basically the circles in the plane, or the spheres in space, etc., and properties are sought which are invariant under various groups of maps on these circles or spheres. Möbius geometry involves those maps which take the set of all circles or spheres through a fixed point into another set of the same sort. Such maps are always induced by similarities and inversions in the plane, space, etc., so that Möbius geometry reduces to the study of this group of maps (the Möbius group) on these spaces. For dimensions $n > 2$ it is thus “conformal geometry”.

In the case of the plane one might also study properties of curves invariant under all analytic maps. I know of only one strange result in this direction—see Theorem 1-4 of Ahlfors {1}.

The more complicated geometries of Laguerre and Lie are allied to the notion of contact transformations (II. (f)). Laguerre geometry involves the maps which take a set of circles tangent to a line (or of spheres tangent to a plane) into another set of this sort, while the geometry of Lie involves the group of maps which simply take circles or spheres which are tangent to each other to pairs of the same sort.

Another weird topic is the theory of webs—see Blaschke {4} and Blaschke and Bol {1}.

All sorts of other oddities may be found by consulting *Mathematical Reviews* and the journal *Tensor*.

(g) *Lie's theory of differential invariants, and É. Cartan's general method of moving frames.* Lie's theory is the one topic which I greatly regret not having written up, for it is used extensively in certain early work which is far more impenetrable than other classical material. In particular, Lie's theory was used in the first investigations of special affine surface theory by Pick [1] and in early work in projective differential geometry (cf. Wilczynski {1}). Unfortunately, a reasonable exposition would probably require close to a hundred pages, which wouldn't fit in anywhere, for the material on first order linear PDE's from Chapter V.10 is needed, while the theory relates most directly to Chapter III.2 (which is already too long). Matters were in no way helped by my lack of understanding, nor, since this delayed its treatment until last, by my lack of endurance.

The most interesting part of Lie's theory applies to situations like Euclidean, special affine, or projective differential geometry, where we seek properties of submanifolds of M which are invariant under a group G of diffeomorphisms [i.e., submanifolds of homogeneous spaces (I. (b))]. The idea is to find “(geometric) differential invariants of order k for r -dimensional submanifolds of M ”, a simple example of which is the curvature κ of a curve c in \mathbb{R}^n . This is an “invariant” (it is the same for a curve and its composition with a Euclidean motion) “of order 2” (one can compute it at any point knowing only the first two derivatives of the curve at that point) which is independent of the parameterization (“geometric”).

Such invariants can be thought of as functions on a suitable space. First we construct the “ k -jets” of maps of $(\mathbb{R}^r, 0)$ into (M, p) ; these are equivalence classes $j_0^k(f)$ of maps $f: (\mathbb{R}^r, 0) \rightarrow (M, p)$, where $f \sim g$ if all mixed partials of order $\leq k$ of all component functions of f and g are equal at 0 (cf. topic VI. (c)). On the set of k -jets represented by immersions $f: (\mathbb{R}^r, 0) \rightarrow (M, p)$

we introduce a further equivalence relation by declaring $j_0^k(f_1)$ and $j_0^k(f_2)$ equivalent if in a neighborhood of $0 \in \mathbb{R}^r$ we have $f_2 = f_1 \circ \alpha$ for some diffeomorphism α . These new equivalence classes are the k^{th} order r -dimensional **contact elements** of M at p (for $k = 1$ they may be identified with the r -dimensional subspaces of M_p). The set $C_r^k M$ of all such contact elements at all $p \in M$ is a manifold, and any diffeomorphism $\phi: M \rightarrow M$ gives rise to a diffeomorphism $\phi_*: C_r^k M \rightarrow C_r^k M$. A geometric differential invariant of order k for r -dimensional submanifolds of M under a group of diffeomorphisms G is just a function $F: C_r^k M \rightarrow \mathbb{R}$ invariant under G (i.e., satisfying $F \circ \phi_* = F$ for all $\phi \in G$). This set-up is briefly discussed in Hermann {2; ch.3, §14}.

We can also consider “differential invariant tensors”. If γ is a k^{th} order r -dimensional contact element of M at p , represented by a k -jet $j_0^k(f)$ for $f: (\mathbb{R}^r, 0) \rightarrow (M, p)$, then the 1st order r -dimensional contact element represented by the 1-jet $j_0^1(f)$ depends only on γ —it may be thought of as the “tangent space” $T\gamma$ of γ . A function F on $C_r^k M$ such that each $F(\gamma)$ is a bilinear function $F(\gamma): T\gamma \times T\gamma \rightarrow \mathbb{R}$, for example, may be thought of as a k^{th} order covariant tensor of order 2 for r -dimensional submanifolds of M ; it is easy to formulate the invariance conditions for such F .

There is actually a reasonable way to compute such differential invariants, or at least to formulate the computations (in practice they become hopeless quite quickly unless one introduces some extraneous geometric insight). Any $X \in \mathfrak{g} =$ Lie algebra of G induces a 1-parameter family of diffeomorphisms $\{\exp tX\}$ of M , hence a family $\{(\exp tX)_*\}$ on $C_r^k M$, and hence a vector field $X^{(k)}$ on $C_r^k M$. If G is connected, and thus generated by the elements $\exp tX_i$ for a basis $\{X_i\}$ of \mathfrak{g} , then F is invariant if and only if $X_i^{(k)}(F) = 0$ for all i ; each of these equations is simply a linear first order PDE in terms of coordinates on $C_r^k M$. This allows one to compute the invariants F once one picks a natural coordinate system on $C_r^k M$ and figures out appropriate methods for evaluating $X_i^{(k)}(\rho)$ for each coordinate function ρ . There is no difficulty solving each particular equation $X_i^{(k)}F = 0$, by the methods of Chapter V.10, §1. We find that F must be constant along certain curves in $C_r^k M$, or equivalently, that F must be expressible as a function of certain combinations of the coordinate functions. The problems arise when we seek a solution F of the equations $X_i^{(k)}F = 0$ for all i ; we then have to guess a single function which can be expressed as a function of each of the different combinations of coordinate functions which arise for each i .

Even without performing the calculations, however, one can decide how many invariants there should be. We seek a submanifold $\mathcal{P} \subset C_r^k M$ which intersects

each “orbit” $\{\phi_*\gamma : \phi \in G\}$, $\gamma \in C_r^k M$ just once. The invariant functions on $C_r^k M$ are in one-one correspondence with the functions on \mathcal{P} , so the number of “independent” ones is the dimension of \mathcal{P} , thus the dimension of $C_r^k M$ minus the dimension of an orbit. Differential invariant tensors can be treated similarly.

Unfortunately, there is no really good reference for this topic. One can try Lie and Scheffers {1; ch.22}, but it will be much easier to read Scheffers {1}. Plane curves are treated in volume 1, part 1, §§8, 9; space curves in volume 1, part 2, §12; and surfaces in volume 2, part 3, §10. All the information here will be quite new, because curves and surfaces are determined up to congruence by certain *functions*, not by tensors; for curve theory this means that no use is made of the parameterization by arclength (which is really equivalent to using the first fundamental form of the image curve). There are also some (rather misleading) calculations in volume 2, part 3, §6 which essentially determine all invariants on the space of jets (i.e., the functions on surfaces which are invariant under Euclidean motions but *not* under change of parameter). The main problem with this reference is that it doesn’t illustrate the general methods of computation outlined above.

The most modern reference for these methods, Guggenheimer {1; ch.7-2}, is frustratingly old-fashioned in its language. It might help to mention that the formally introduced new “variables” $x'_i = dx_i/dx_1, x''_i, \dots$ are merely certain natural coordinates for 1-dimensional contact elements in \mathbb{R}^m : on the set of 1-dimensional contact elements γ represented by curves whose images can be parameterized as $x_1 \mapsto (x_1, f_2(x_1), \dots, f_m(x_1))$ [for unique f_2, \dots, f_m] we let $x'_i(\gamma) = f'_i$, etc. Similarly, on the set of 2-dimensional contact elements γ in \mathbb{R}^3 which are represented by immersions of \mathbb{R}^2 whose images can be parameterized as $(x_1, x_2) \mapsto (x_1, x_2, f(x_1, x_2))$ we have the coordinates $p(\gamma) = \partial f / \partial x, q(\gamma) = \partial f / \partial y, r(\gamma) = \partial^2 f / \partial x^2$, etc. Beware of misprints in some of the computations of the “prolongations”.

There are extensive calculations in Forsythe {1; §§132–146}, but they are not carried out directly on the contact elements. Instead, “relative invariants of order w ” are computed first; these are functions F on the jets such that

$$F(j_0^k(f \circ \alpha)) = (\det \alpha)^w F(j_0^k(f))$$

for any diffeomorphism α . Clearly the product of relative invariants whose weights add up to 0 will give invariants on the contact elements. (Similar tricks are used in the works of Pick and Wilczynski quoted above, as well as in the much more straightforward calculations in Blaschke {1; v.3}.) One may also try the introductory chapter of Schirokow and Schirokow {1}.

Lie's methods are supposedly applicable, in ways I don't understand at all, to such problems as the equivalence of Riemannian manifolds, and even more general questions. See Lie and Scheffers {1; ch.23}, Veblen {1; ch.3, §§20–22, ch.5, 6}, and Wright {1}. It might also be mentioned that there is a theory involving even more general notions than jets and contact elements, the “geometric objects”—see Yano {3; ch.2} or Aczél and Gołab {1}.

É. Cartan's general method of moving frames is a sort of dual to Lie's method which allows computations to be made more easily. The general features of the theory are well illustrated by the development of special affine surface theory in Chapter III.2—see especially pg. III.102. We consider moving frames X_1, X_2, X_3 along $M^2 \subset \mathbb{R}^3$ with $\det(X_1, X_2, X_3) = 1$. An arbitrary moving frame of this sort is what Cartan calls a “zeroth order frame”. A “first order frame” is one which is adapted to M (i.e., X_1, X_2 are tangent to M). To define “second order frames” we now try to specialize the first order frames as much as possible by seeking an appropriate condition on the dual and connection forms. As we observed on pg. III.82ff., the condition $\psi_i^3 = \theta^i$ has just the “invariant” property we need—it depends only on the value of the frame at p [i.e., if X_1, X_2, X_3 and $\bar{X}_1, \bar{X}_2, \bar{X}_3$ are adapted moving frames with dual and connection forms θ^i, ψ_j^i and $\bar{\theta}^i, \bar{\psi}_j^i$, respectively, and the two frames agree at p , then $\psi_i^3(p) = \theta^i(p)$ if and only if $\bar{\psi}_i^3(p) = \bar{\theta}^i(p)$].

This definition of second order frames already gives an invariant tensor, the special affine first fundamental form. To obtain this specialization we used the calculations on pp.III.79–81. To be sure, these calculations were not very difficult, but that is because we already knew what we were looking for; we didn't even bother to compute the ψ_j^i in general, since they would not be involved in our invariant condition. It would have been much harder to simply guess an invariant condition without the previous geometric motivation (indeed, the difficulties which arise here are exactly dual to the problem in Lie's method of finding a common solution to the equations $X_i^{(k)} F = 0$); this is an aspect of the theory which Cartan always carefully suppressed in his expositions of it.

Now we can seek “third order frames” by specializing the second order frames. An appropriate condition is $\psi_3^3 = 0$; it gives us the special affine normal. The verification that this condition is invariant is given on pp. III.102–103. It would clearly be a lot easier to discover *ab initio* than the corresponding invariant condition (pg. III.100) for first order frames; in general, one always works with the highest order frames already successfully discovered. Specializing the third order frames would lead us to the special affine second fundamental form (X_1 and X_2 would be the eigenvectors of $\mathbf{\Pi}$ with respect to \mathbf{I}). For more examples and details, see Cartan {6} or Favard {1}.

IV. The Russian school; synthetic differential geometry

A thorough treatment of the foundations of surface theory without differentiability hypotheses is given in Alexandrov {1}. As an introduction, the reader may prefer to consult Busemann {1} or relevant portions of Efimov {1} and Pogorelov {1}, {2}, {3}. For an extensive connected account of further developments in the theory see Pogorelov {5}. We might also mention Pogorelov {6}, where the geometric results are used to obtain stronger-than-usual results about PDE's.

Although most of the material in these references pertains to convex surfaces (for which one may also consult Bonneson and Fenchel {1}, Blaschke {2}, Hadwiger {1}, and Yaglom and Boltyanskii {1}), there is also an elaborate theory which investigates the most general sorts of surfaces, or even arbitrary metric spaces. One treatment can be found in Alexandrov and Zalgaller {1}, while a somewhat different direction is taken in Blumenthal {1}, Blumenthal and Menger {1}, and Rinow {1}. In yet another vein, we have the work of Busemann, which represents the very antithesis of Riemann's approach (whereby a mechanism for measuring lengths of arbitrary curves is postulated, and geodesics are defined as curves of minimal length). In Busemann {3} postulates are instead given for the geodesics, and many relations of classical differential geometry are derived from them; although the arguments are often involved, just as a rigorous axiomatic development of Euclidean geometry would be, the strength of the results is often startling. For related results see Busemann {2}.

V. Applications to physics

(a) *Classical mechanics*. It turns out that differential geometry provides the natural language for classical mechanics, for the two equivalent basic formulations of the subject, via Lagrange's equations and Hamilton's equations, take place on the tangent bundle and cotangent bundle, respectively, of a suitable manifold. A discussion of mechanics which manages to make some interesting points in a short space, but which doesn't make very clear which manifold one is working on, may be found in Laugwitz {1: §14}. Another brief discussion, without this shortcoming, is contained in Bishop and Goldberg {1: ch.6}. One can consult pp. 141–147 of Sternberg {1}, which also mentions other aspects of the subject as part of an extensive discussion of the calculus of variations, in the succeeding chapter. Similarly, see Hermann {1: ch.16}. The short book by Godbillon {1} reaches its climax in the final, 9 page, chapter on mechanics, which begins by defining a "mechanical system" as a triple (M, T, π) , where M is a manifold, T a differentiable function on TM , and π a semi-basic form on TM . A serious study of mechanics will be found in Abraham {1}, beginning in the third chap-

ter, where it is admitted however, that the treatment “possibly . . . will seem severely unmotivated without some background in classical mechanics . . .”; see also Wasserman {1}. Aside from this difficulty, common to all these books, the most heartbreaking omission is any adequate discussion of the fictional “forces of constraint” which are involved in such useful abstractions as “rigid bodies”. These can usually be formulated as particular subspaces that the velocity vector in \mathbb{R}^{3N} of the system of N particles is constrained to lie in. When these subspaces form an integrable distribution (the constraints are “holonomic”) the integral submanifolds form a lower dimensional “configuration space” and the basic principle is that the motion of the system with these forces of constraint is determined by restricting the original equations on $T\mathbb{R}^{3N}$ or $T^*\mathbb{R}^{3N}$ to equations on the tangent or cotangent bundle of the configuration space. (Some details are given in Hermann {2; ch.2}.) For physics books on classical mechanics the following references may be of use: Corbin and Stehle {1} (the most modern in spirit of the elementary books), Pars {1}, and Whittaker {1}.

(b) *General Relativity*. Fortunately, I was ultimately not foolish enough to attempt writing anything on this vast subject which I do not understand. My conscience is set at ease by recommending the monumental book of Misner, Thorne, and Wheeler {1}, which is probably owned by every relativist in the world. For more mathematical treatments you may prefer Sacks and Wu {1}, which places great emphasis on foundational points, Weinberg {1}, Hawking and Ellis {1}, O’Neill {2}, and Beem {1}.

VI. Miscellaneous

(a) *Calculus of variations; Hamilton-Jacobi theory*. A subject closely linked with mechanics (V.(a)). The two great classical works usually referred to are Carathéodory {1} and É. Cartan {4}. See also Abraham {1}, H. Cartan {1}, Godbillion {1}, Hermann {1; pt.2}, Rund {2}, and Sternberg {1; ch.4}.

(b) *Sprays*. This is a topic which I have assiduously avoided learning, convinced that one can get by without it, and suspicious that it’s just a complicated new way of saying something old. Less obstinate readers may wish to consult Gromoll, Klingenberg, and Meyer {1; 60}, Lang {1; ch.4, §§3–5, ch.6, §6}, or Sternberg {1; 199, 361}.

(c) *Jets*. These, and the contact elements [cf. III.(g)], are natural structures to consider in differential geometry, but they are only just beginning to be used in any serious way. For basic definitions, see Bourbaki {1; §12} or Dieudonné {1; v:3 (ch.16.5–Problem 9, ch.16.9–Problem 1), v:4 (ch.20.1–Problem 3)}. For some applications, see Kobayashi {1; 139ff.}.

(d) *Other definitions of connections.* Gromoll, Klingenberg, and Meyer {1; 43} gives a definition of a connection in terms of a map $K: TTM \rightarrow TM$. This is useful in dealing with infinite dimensional manifolds; see Flaschel and Klingenberg {1; ch.1}. Connections have also been defined as sprays with certain properties, and as a splitting of the “jet exact sequence”. I personally feel that the next person to propose a new definition of a connection should be summarily executed.

(e) *Reducing differentiability hypotheses.* It is of some interest to some people (analysts) to establish results with the minimum differentiability assumptions. Although the work of the Russian school sometimes eclipses such efforts, this is not always true, and in any case few mathematicians seem to find it a sufficiently compelling argument to go that route. In any proof of classical geometry one can always carefully count how many times one differentiates, but this usually turns out to be one or two more times than one really has to, if one is sufficiently clever. So the problem of finding minimal differentiability hypotheses (and examples to prove they are minimal) is not easy. A long series of papers on this subject was published by Hartmann and Wintner, mainly in the *American Journal of Mathematics*, beginning in 1947.

(f) *Transversality.* Not a part of differential geometry, really, but of differential topology. Nevertheless, it is probably a wise move to learn the basic ideas. See Sternberg {1; 64ff.} or Guillemin and Pollack {1}, with its many beautiful applications.

(g) *Polyhedral geometry, models, constructive aspects, etc.* Perhaps the oldest contribution in this direction was the argument of Hilbert and Cohn-Vossen {1; pp. 194–195} proving Gauss’ Theorema Egregium for polyhedral surfaces. Recent work of T. Banchoff and others has carried this approach much further. One may also consult Sauer {1} and Kruppa {1}. For a rather different approach to the Gauss-Bonnet-Chern Theorem, see Palais {1}.

B. BOOKS

During the compilation of this bibliography, certain instincts urged me to seek encyclopaedic completeness, while healthier ones advised selectivity and utility. From this conflict resulted the usual unsatisfactory compromise, wherein the advantages of neither course of action is retained. I have tried to single out sources which might be particularly valuable, but this applies mainly to books concerned with the topics covered in these five volumes; many others will have already been mentioned in Part A.

Encyklopädia der Mathematischen Wissenschaften, Volume III, Part 3D, B. G. Teubner, Leipzig, 1902–1927.

Abraham, R. and Marsden, J.

- {1} *Foundations of mechanics*, 2nd ed., Addison-Wesley, Reading, Mass., 1978 (MR 81e:58025).

Aczél, J. and Gołab, S.

- {1} *Funktionalgleichungen der Theorie der Geometrischen Objekte*, Państwowe Wydawnictwo Naukowe, Warsaw, 1960 (MR 24 #A3588).

Ahlfors, L. V.

- {1} *Conformal Invariants: topics in geometric function theory*, McGraw-Hill, New York, 1973 (MR 50 #10211).

Akivis, M. A. and Goldberg, V. V.

- {1} *Projective Differential Geometry of Submanifolds*, North-Holland, Amsterdam, 1993 (MR 94i:53001)

Alexandrov, A. D.

- {1} *Die Innere Geometrie der Konvexen Flächen*, Akademie-Verlag, Berlin, 1955 (MR 17 #74).
- {2} *Kurven und Flächen*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1959 (MR 21 #3866).

A very nice elementary introduction to curves and surfaces which mentions some things you might not see otherwise (e.g., why a pail with a curved rim is stronger than one with a plain rim). For an English translation see Chapter 7 of

Alexandrov, A. D., Kolmogorov, A. N., and Lavrent'ev, M. A. (eds.)

- {1} *Mathematics. Its contents, methods and meaning*, 2nd ed., M.I.T. Press, Cambridge (Mass.), 1969 (MR 39 #1258a–c).

Alexandrov, A. D. and Zalgaller, V. A.

- {1} *Intrinsic Geometry of Surfaces*, American Mathematical Society, Providence, R.I., 1967 (MR 35 #7267).

Arnold, V. I.

- {1} *Mathematical Methods in Classical Mechanics*, Springer-Verlag, New York, 1978 (MR 57 #14033b).

Auslander, L.

- {1} *Differential Geometry*, Harper & Row, New York, 1967 (MR 35, #2208).

This is an attempt to construct an introductory course in differential geometry from the point of view of Lie groups, with the fundamental equations of surface theory arising from the equations of structure of $SO(3)$. Later chapters cover Riemannian geometry. The treatment of geodesic completeness (pp. 203–214) may be of interest, and the Poincaré upper half-space is discussed in some detail (pp. 223–236). In particular, there is a description of the various one-parameter subgroups of the group of isometries. The orbits of these subgroups are the geodesics, geodesic circles, horocycles, and equidistant curves (for these are the curves of constant curvature).

Auslander, L. and MacKenzie, R. E.

- {1} *Introduction to Differentiable Manifolds*, McGraw-Hill, New York, 1963 (MR 28 #4462).

Beem, J. K., Ehrlich, P. E., and Easley, K. L.

- {1} *Global Lorentzian Geometry*, 2nd ed., Marcel Dekker, Inc., New York, 1996 (MR 97f:53100).

Berger, M.

- {1} *Lectures on Geodesics in Riemannian Geometry*, Tata Institute of Fundamental Research, Bombay, 1965 (MR 35 #6100).

These notes cover many topics, frequently with details not to be found elsewhere.

see also Lascoux, A.

Berger, M., Gauduchon, P., and Mazet, E.

- {1} *Le Spectre d'une Variété Riemannienne*, Springer-Verlag, Berlin, 1971 (MR 43 #8025).

Berger, M. and Gostiaux, B.

- {1} *Géométrie Différentielle*, Armand Colin, Paris, 1972 (MR 58 #13102)

A very beautiful recent text on differentiable manifolds, with some differential geometry included. §§6.7–6.9 and 7.5 prove the Gauss-Bonnet

theorem for submanifolds of Euclidean space by the method of Allendoerfer and Fenchel mentioned at the beginning of Chapter 13 (pg. V.264). Chapter 9 treats global properties of curves, including an elementary proof of the Jordan curve theorem for smooth curves, Whitney's theorem on smooth homotopy of closed curves, and the formula of Fabricius-Bjerre-Halpern, which relates the number of double points and inflection points of a closed curve to the number of double tangents (lines tangent to the curve at two different points).

The next reference is a new edition, with additional material.

- {2} *Géométrie Différentielle: variétés, courbes et surfaces*, Presses Universitaires de France, Paris, 1987 (MR 89b:53001).

There is also an English translation:

Differential Geometry: manifolds, curves, and surfaces, Springer-Verlag, New York-Boston, 1988 (MR 88h:53001).

Besse, A. L.

- {1} *Manifolds All of Whose Geodesics are Closed*, Springer-Verlag, New York, 1978 (MR 80c:53044).

Bianchi, L.

- {1} *Vorlesungen über Differentialgeometrie*, B. G. Teubner, Leipzig, 1899.

This is a translation, with some additions, of the first edition of the original Italian work. Naturally it contains a considerable number of results from classical surface theory, but it differs from many classical books by also treating surfaces in the spaces of constant curvature. See in particular §348, which considers a surface M in the upper half-space model of H^3 . For $p \in (x, y)$ -plane, let γ be the geodesic intersecting M orthogonally and approaching p , and let $f(p)$ be the other point in the (x, y) -plane which γ approaches. Then f is holomorphic if and only if M has (intrinsic) curvature 0. Also note that §110 gives a nicer treatment of the problem considered on pp. V.217–218; Bianchi shows directly that the curve $t \mapsto (u(t, 0), v(t, 0), w(t, 0))$ has the same curvature and torsion as c . (The whole problem is simply ignored by Darboux {1}. By the way, there is no adequate treatment anywhere in the classical literature of the case where $\kappa = \tilde{\kappa}_g$.)

There is a later, 4 volume, Italian edition, not translated, which treats special questions of surface theory in great detail.

Bishop, R. L. and Crittenden, R. J.

{1} *Geometry of Manifolds*, Academic Press, New York, 1964 (MR 29 #6401).

Very compactly written, with many results merely quoted or left as exercises. Particular attention may be called to some of the problems on pp. 106–107, 110, 114, 134. There is a study of complete simply-connected manifolds of constant curvature (§9.5) rather different from the elementary one outlined in Problem III.1-5, a more elaborate study of convex neighborhoods, using the second variation (§11.8), and some applications of the second variation to theorems on the volumes of balls (pp. 256–257). A bibliography of 95 items.

Bishop, R. L. and Goldberg, S. I.

{1} *Tensor Analysis on Manifolds*, Macmillan, New York, 1968 (MR 36 #7057).

Blaschke, W.

{1} *Differential Geometrie*, Volumes 1, 2, Chelsea, New York, 1967; Volume 3, Springer, Berlin, 1929.

Although this book is quite old-fashioned, I nevertheless find it very stimulating, perhaps because the author is more interested in genuine geometric questions, especially global ones, than in the formalities of calculations. More topics are covered here than in almost any other classical book, and there is an extraordinary number of interesting exercises, remarks, and sidelights.

Volume 1, §72 shows that if the geodesic circles are the same as the curves of constant κ_g , then K is constant, while §84 proves the more difficult result (mentioned on pg. IV.309) that K is constant if all curves of constant κ_g are closed. The manipulations of §94 (used in the next section for a proof of Christoffel's theorem) are mysterious; Problem III.3-8 may be used instead. See also the funny result on pg. 121. It is interesting to find that the general formula for variation of area was already given by Gauss (§109). Classical results of Schwarz (one mentioned under topic II.(a)) are given in §§110, 111, while §116 gives the second variation of area (in a special case), and mentions a classical condition of Schwarz for a minimal surface to be a local minimum for area. For a modern presentation, see Barbosa and do Carmo [1]. The proof of the related Theorem IV.9-39 is from Rado {1}. §117 gives the first variation of H and K (we essentially found the first variation of H in order to find the second variation of area). Problem 2 in §118 mentions interesting properties of associated minimal surfaces, for example, the tangent planes at corresponding points are parallel. Conversely, if there is an isometry between two surfaces such that tangent planes are parallel, then they are

either congruent surfaces or associated minimal surfaces (to be taken with a grain of salt—one of the surfaces could be the union, along a common curve, of a piece congruent to part of the other surface and another piece associated to a part of the other surface which is a minimal surface).

Volume 2 covers affine differential geometry. There are very many nice geometric interpretations of the invariants arising here, as well as many global results.

For Volume 3 see topic III.(f).

- {2} *Kreis und Kugel*, de Gruyter & Co., Berlin, 1956 (MR 17 1123).
- {3} *Vorlesungen über Integralgeometrie*, 2nd ed., Chelsea, New York, 1949.
- {4} *Einführung in die Geometrie der Waben*, Birkhäuser, Basel, 1955 (MR 17 780).

Blaschke, W. and Bol, G.

- {1} *Geometrie der Gewebe*, Springer, Berlin, 1938.

Blaschke, W. and Leichtweiss, K.

- {1} *Elementare Differentialgeometrie*, 5th ed., Springer-Verlag, Berlin, 1973 (MR 50 #3122).

This is a modernization of Blaschke's book which preserves the style of the original. Numerous new problems and references of interest.

Blaschke, W. and Reichardt, H.

- {1} *Einführung in die Differentialgeometrie*, 2nd ed., Springer-Verlag, Berlin, 1960 (MR 22 #7062).

This is an attempt to modernize Blaschke's book by writing everything in terms of moving frames. It may be consulted for a few interesting points hard to find elsewhere, especially §§56, 57, 77, and 69, Problem 19.

Blumenthal, L. M.

- {1} *Theory and Applications of Distance Geometry*, 2nd ed., Chelsea, New York, 1970 (MR 42 #3678).

Blumenthal, L. M. and Menger, K.

- {1} *Studies in Geometry*, W. H. Freeman, San Francisco, 1970 (MR 42 #8370).

Bochner, S.

see Yano, K.

Bol, G.

- {1} *Projektive Differentialgeometrie*, 3 vols., Vandenhoeck & Ruprecht, Göttingen, 1950 (MR 16 1150).

Extensive bibliography, extending that of Fubini and Čech {1}.
see also Blaschke, W.

Boltyanskii, V. G.

- {1} *Envelopes*, Pergamon Press, Macmillan, New York, 1964 (MR 31 #2348).
see also Yaglom, I. M.

Bonneson, T. and Fenchel, W.

- {1} *Theorie der Konvexen Körper*, Springer, Berlin, 1934.

Boothby, W.

- {1} *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975 (MR 54 #13956).

Boothby, W. and Weiss, G. L. (eds.)

- {1} *Symmetric Spaces: Short courses presented at Washington University*, Dekker, New York, 1972 (MR 53 #687).

Bourbaki, N.

- {1} *Variétés Différentielles et Analytiques. Fascicule de Résultats/Paragaphes 1 à 7 and /Paragaphes 8 à 15*, Hermann, Paris, 1971 (MR 43 #6834).

Bourbaki is the originator of that famous pedagogical method whereby one begins with the general and proceeds to the particular only after the student is too confused to understand even that any more. His influence is to be seen everywhere, probably in these volumes too. Bourbaki has apparently decided that the theory of manifolds has now entered that domain of “dead” mathematics to which he hopes to give definitive form. In this summary of results the corpse is laid out to public view; the complete autopsy is eagerly awaited.

Boys, C. V.

- {1} *Soap Bubbles, their colors and the forces which mold them*, 3rd ed., Dover, New York, 1959.

Brakke, K. A.

- {1} *The Motion of a Surface by its Mean Curvature*, Princeton University Press, Princeton, N.J., 1978 (MR 82c:49035).

Brickell, F. and Clark, R. S.

- {1} *Differentiable Manifolds*, Van Nostrand Reinhold, London, 1970.

Bryant, R. L., Chern, S.-S., Gardner, R. B., Goldschmidt, H. L., and Griffiths, P. A.

- {1} *Exterior differential systems*, Mathematical Sciences Research Institute Publications, 18, Springer-Verlag, New York, 1991 (MR 92h:58007).

This very valuable book is concerned with an important, but extremely difficult, portion of É. Cartan’s work. There is an extended discussion

of matters related to differential systems (Chapter 10, Addendum 1), the Cartan-Janet theorem (Chapter 11, Addendum), and many related topics.

Busemann, H.

- {1} *Convex Surfaces*, Interscience, New York, 1958 (MR 21 #3900).
- {2} *Recent Synthetic Differential Geometry*, Springer-Verlag, Berlin, 1970 (MR 45 #5936).
- {3} *The Geometry of Geodesics*, Academic Press, New York, 1955 (MR 17 779).

Campbell, J. E.

- {1} *A Course of Differential Geometry*, Oxford University Press, Oxford, 1926.

§§149–154 prove that any n -dimensional Riemannian manifold can be imbedded in an $(n + 1)$ -dimensional Einstein space of vanishing scalar curvature. I know of no other reference for this result.

Carathéodory, C.

- {1} *Calculus of Variations and Partial Differential Equations of the First Order*, Holden-Day, San Francisco, 1965 (MR 33 #597).

Carmo, M. do

- {1} *Differential Geometry of Curves and Surfaces*, Prentice Hall, Englewood Cliffs, N.J., 1976 (MR 52 #15253).
- {2} *Notas de Geometria Riemanniana*, Instituto de Matematica Pura e Aplicada, Rio de Janeiro, 1972.

Carrell, J. B.

see Dieudonné, J. A.

Cartan, É.

- {1} *Oeuvres Complètes*, 3 vols., in 6 parts, Gauthier-Villars, Paris, 1952–1955 (MR 14 343; 15 383; 16 697).

The greatest differential geometer of the previous generation. Few have read his works, many pretend to have read them, and every one agrees that every one should read them. I get shell-shock every time I try. Fortunately, most of his books have by now been reworked in modern presentations, and it's still worth looking at the originals after you know more or less what they're about.

- {2} *Les Systèmes Différentiels Extérieurs et leurs Applications Géométriques*, Hermann, Paris, 1971 (MR 7 520d).

As an introduction, try Godbillion {1}.

- {3} *Leçons sur la Géométrie des Espaces de Riemann*, 2nd. ed., Gauthier-Villars, Paris, 1963 (MR 8 602g).

This is probably the easiest and most important of Cartan's books. Most of the material has been covered somewhere in these five volumes. There are few other classical books with a thorough description of the spaces of constant curvature—unfortunately, here it is approached via projective geometry.

Also available in English translation, with additional notes:

Geometry of Riemannian Spaces, Math Sci Press, Brookline, Mass., 1983 (MR 85m:53001)

- {4} *Leçons sur les Invariants Intégraux*, Hermann, Paris, 1971 (MR 50 #8238).

The hardest part of this book is the Cartan-Kähler theorem. For further reading in modern sources, see Dieudonné {1; v.4, ch.18, §§8–18}, the very useful discussion in Jacobowitz and Moore [1], and Bryant {1}. The second half of the book gives applications to all sorts of questions in differential geometry, mainly from surface theory.

- {5} *Leçons sur la Théorie des Espaces à Connexion Projective*, Gauthier-Villars, Paris, 1937.

- {6} *La Théorie des Groupes Finis et Continus et la Géométrie Différentielle traitées par la Méthode du Repère Mobile*, Gauthier-Villar, Paris, 1937.

Cartan, H.

- {1} *Formes Différentielles*, Hermann, Paris, 1967 (MR 37 #6858).

A neat little presentation of the elements of manifold theory, and the calculus of variations. In the midst of all this elegance the proof on pg. 162 is truly startling.

Čech, E.

see Fubini, G.

Cecil, T.

- {1} *Lie Sphere Geometry*, Springer-Verlag, New York, 1992 (MR 94m:53076).

Chavel, I.

- {1} *Riemannian Symmetric Spaces of Rank One*, M. Dekker, New York, 1972 (MR 52 #4185).

Discusses a class of spaces connected with the sphere theorem, when the curvature is allowed to take on values K with $1/4 \leq K \leq 1$.

{2} *Riemannian Geometry. A modern introduction*, Cambridge University Press, Cambridge, 1993 (MR 95j:53001).

{3} *Eigenvalues in Riemannian Geometry*, Academic Press, Orlando, Fla., 1984 (MR 86g:58140).

Cheeger, J. and Ebin, D.

{1} *Comparison Theorems in Riemannian Geometry*, North-Holland, Amsterdam, 1974 (MR 56 #16538).

Chen, B.-Y.

{1} *Geometry of Submanifolds*, Dekker, New York, 1973 (MR 50 #5697).

Many results, using various techniques, about submanifolds (especially in the spaces of constant curvature), involving many of the topics in Volumes III–V. Extensive bibliography.

{2} *Geometry of Submanifolds and its Applications*, Science University of Tokyo, Tokyo, 1981 (MR 82m:53051).

Chern, S.-S.

{1} *Complex Manifolds without Potential Theory*, Van Nostrand, Princeton, N.J., 1967 (MR 37 #940).

{2} *Topics in Differential Geometry* (mimeographed notes), The Institute for Advanced Study, Princeton, N.J., 1951 (MR 19 764).

{3} (ed.) *Studies in Global Geometry and Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1967 (MR 35 #1429).

The next reference is a later, expanded, version:

{4} (ed.) *Global Differential Geometry*, Mathematical Association of America, Washington, DC, 1989 (MR 90d:53003).

see also Bryant, R. L.

Chernov, I.

{1} *1,000 Best Short Games of Chess*, Simon and Schuster, New York, 1955.

Choquet-Bruhat, Y.

{1} *Géométrie Différentielle et Systèmes Extérieurs*, Dunod, Paris, 1968 (MR 38 #5118).

Clark, R. S.

see Brickell, F.

Cohn-Vossen, S.

see Hilbert, D.

Coolidge, J. L.

- {1} *A History of Geometrical Methods*, Dover, New York, 1963 (MR 28 #3357).

I can't understand any of the mathematics in this book, but there must be something to be learned from a man who can say "as Darboux once remarked to me ...".

Corbin, H. and Stehle, P.

- {1} *Classical Mechanics*, 2nd ed., Wiley, New York, 1960 (MR 22 #3131).

Courant, R.

- {1} *Dirichlet's Principle, Conformal Mapping and Minimal Surfaces*, Interscience, New York, 1950 (MR 12 90).

Courant, R. and Hilbert, D.

- {1} *Methods of Mathematical Physics*, vol. 2, Interscience, New York, 1962 (MR 25 #4216).

Crittenden, R. J.

see Bishop, R. L.

Darboux, G.

- {1} *Leçons sur la Théorie Générale des Surfaces*, 4 vols., 3rd ed., Chelsea, New York, 1972.

The great classic. Always referred to in hushed tones of awe. Unreadable, but in this new printing by Chelsea, at just \$60 for 2200 pages, how can you resist?

Reprinted in 1993 by Éditions Jacques Gabay, Sceaux.

- {2} *Leçons sur les Systèmes Orthogonaux et les Coordonnées Curvilignes*, Gauthier-Villars, Paris, 1898.

de Rham, G.

see Rham, G. de

Dieudonné, J. A.

- {1} *Éléments d'Analyse*, volumes 3, 4, Gauthier-Villars, Paris, 1971 (MR 42 #5266).

Dieudonné, J. A. and Carrell, J. B.

- {1} *Invariant Theory, Old and New*, Academic Press, New York, 1971 (MR 42 #4828).

do Carmo, M.

see Carmo, M. do

Dodson, C. T. J. and Poston, T.

- {1} *Tensor Geometry. The geometric viewpoint and its uses*, Pitman, London, 1977 (MR 58 #18158).

Particular attention is given to indefinite Riemannian metrics.

Easley, K. L.

see Beem, J. K.

Eberlein, P.

- {1} *Geometry of Nonpositively Curved Manifolds*, University of Chicago Press, Chicago, Ill., 1996 (MR 98h:53002).

Ebin, D.

see Cheeger, J.

Efimov, N. W.

- {1} *Flächenverbiegung im Grossen*, mit einem Nachtrag von E. Rembs und K. P. Grotemeyer, Akademie-Verlag, Berlin, 1957 (MR 19 59).

A wonderful review article, covering many different aspects of rigidity, with additions by E. Rembs and K. P. Grotemeyer on more recent (but usually rather specialized) results which just double the size of the original. In addition to providing an introduction to the methods of A. D. Alexandrov, there are descriptions of the work of H. Lewy, Liebmann, Cohn-Vossen, etc., which cannot be found collected together anywhere else. I'm afraid, however, that the purported proof of Satz X on pg. 177 is invalid—the mistake appears at the very last step; only Pogorelov's difficult proof (for the more general situation) is available for this result.

The original article also appears in the A.M.S. Translations, series 1, volume 6.

Ehrlich, P. E.

see Beem, J. K.

Eisenhart, L. P.

- {1} *Riemannian Geometry*. Princeton University Press. Princeton, N.J., 1966 (MR 11 687g).

As good a compendium as any of the basic material of Riemannian geometry. In addition to isometric correspondence, conformal correspondence is also allotted an important role. (However, the discussion of the conformal tensor (§28) is treated better in Gerretsen {1}.) The important classical notion of "scalar curvature" is defined on pg. 83. Appendix 3 gives the form of the metric in normal coordinates, and Appendix 22

gives the surprising result that the Codazzi-Mainardi equations follow from the Gauss equations for “general” hypersurfaces of \mathbb{R}^n , $n \geq 4$. Apart from this, you might like to look at the theorems on pp. 124, 155, 179, 182, 183, 184, 199.

- {2} *A Treatise on the Differential Geometry of Curves and Surfaces*, Dover, New York, 1960 (MR 22 #5936).

This is a subset of Darboux, though it uses classical calculations rather than the moving frames which Darboux introduced for surface theory. A good way to get into more esoteric aspects of surface theory. Many results given in the problems.

- {3} *Non-Riemannian Geometry*, American Mathematical Society, New York, 1927 (MR 98j:53001).

I.e., the study of general connections.

- {4} *Continuous Groups of Transformations*, Princeton University Press, Princeton, N.J., 1933.

- {5} *Transformations of Surfaces*, 2nd ed., Chelsea, New York, 1962.

A complement to Eisenhart {2}, going into greater detail on various classical transformations of surfaces. It doesn't look very interesting, but parts may turn out to be important, as they involve contact transformations (topic II.(f)).

- {6} *An Introduction to Differential Geometry, with use of the tensor calculus*, Princeton University Press, Princeton, N.J., 1940 (MR 2 154).

Ellis, G. F. R.

see Hawking, S. W.

Favard, J.

- {1} *Cours de Géométrie Différentielle Locale*, Gauthier-Villars, Paris, 1957 (MR 18 668).

A complete development of differential geometry, covering many topics in considerable detail, using É. Cartan's general method of moving frames (topic III.(g)), which is expounded at the beginning, before any geometry has been done. Shades of Bourbaki!

Fenchel, W.

see Bonnesen, T.

Ferus, D.

- {1} *Totale Absolutkrümmung in Differentialgeometrie und -topologie*, Springer-Verlag, Berlin, 1968 (MR 40 #3468).

Finsler, P.

- {1} *Über Kurven und Flächen in Allgemeinen Räumen*, Birkhäuser, Basel, 1951 (MR 13 74).

Flanders, H.

- {1} *Differential Forms*, Academic Press, New York, 1963 (MR 28 #5397).

Flaschel, P. and Klingenberg, W.

- {1} *Riemannsche Hilbertmannigfaltigkeiten. Periodische Geodätsche*, Springer-Verlag, Berlin, 1972 (MR 49 #6275).

Forbes, W. F.

see Rund, H.

Forsyth, A. R.

- {1} *Lectures on the Differential Geometry of Curves and Surfaces*, Cambridge University Press, Cambridge, 1912.

Like Eisenhart {2}, a subset of Darboux, and a good way to get into surface theory. §§132–146 will probably be incomprehensible (cf. topic III.(g)).

Fubini, G. and Čech, E.

- {1} *Introduction a la Géométrie Projective Différentielle des Surfaces*, Gauthier-Villars, Paris, 1931.

Large bibliography.

Gardner, R. B.

see Bryant, R. L.

Gauduchon, P.

see Berger, M.

Gerretsen, J. C. H.

- {1} *Tensor Calculus and Differential Geometry*, P. Noordhoff N.V., Groningen, 1962 (MR 25 #1494).

An introduction to Riemannian geometry about midway between Eisenhart's presentation and a completely modern one. Considerably less material than in Eisenhart, but there is a discussion of integrability conditions, which Eisenhart always treats rather shabbily, so it is a preferable source for certain results, notably the Weyl-Schouten conditions for a conformally flat space (pg. 188).

Gilkey, P. B.

- {1} *The Heat Equation, The Index Theorem, and the Atiyah-Singer Index Theorem*, 2nd ed., CRC Press, Boca Raton, FL, 1995 (MR 98b:58156).

Includes an introduction to pseudo-differential operators, which are used to obtain both Hodge's theorem and the Gauss-Bonnet-Chern theorem.

Godbillon, C.

- {1} *Géométrie Différentielle et Mécanique Analytique*, Hermann, Paris, 1969 (MR 39 #3416).

Goetz, A.

- {1} *Introduction to Differential Geometry*, Addison-Wesley, Reading, Mass., 1968 (MR 42 #2370).

Gołab, S.

see Aczél, J.

Goldberg, S. I.

- {1} *Curvature and Homology*, Academic Press, New York, 1962 (MR 25 #2537).

Basically a compendium of results.

see also Bishop, R. L. and Weber, W. C.

Goldberg, V. V.

see Akivis, M. A.

Goldschmidt, H. L.

see Bryant, R. L.

Gostiaux, B.

see Berger, M.

Goursat, É.

- {1} *Cours D'Analyse Mathématique*, 5th ed., vol. 1, Gauthier-Villars, Paris, 1933.

The exercises for chapter 12 give a classical style proof of Laguerre's theorem (pg. III.193), as well as the result of Beltrami given in Problem III.4-4.

Gray, A.

- {1} *Modern differential geometry of curves and surfaces*, CRC Press, Boca Raton, Fla., 1993 (MR 95g:53002).

Greub, W., Halperin, S., and Vanstone, R.

- {1} *Connections, Curvature, and Cohomology*, 3 vols., Academic Press, New York, 1972, 1973, 1976 (MR 48 #1423, #1424; 53 #4110).

A very thorough treatise on the subjects indicated by the title, with information on an extremely large number of topics. Rather heavy on

the symbolism, which is frequently neither standard nor felicitous. Large bibliography, especially for volume 2.

Griffiths, P.

- {1} *Topics in Algebraic and Analytic Geometry*, Princeton University Press, Princeton, N.J., 1974 (MR 50 #7596).

see also Bryant, R. L.

Gromoll, D., Klingenberg, W., and Meyer, W.

- {1} *Riemannsche Geometrie im Grossen*, Springer-Verlag, Berlin, 1968 (MR 37 #4751).

These lecture notes, by three important differential geometers, give a modern presentation of intrinsic Riemannian geometry that starts right at the beginning and gets into recent deep global results. In addition, there are hundreds of examples and exercises that significantly extend the theory. By now many of the proofs have undoubtedly been simplified, and eventually the notes may become outdated, especially in light of Cheeger and Ebin {1}, but they are definitely worth owning. We cannot mention here all the topics covered, but the Sphere Theorem is certainly one of the most significant. There is a version of Klingenberg's Theorem (Theorem IV.8-36) for arbitrary dimensions (pg. 254), and attention might also be called to the Lemma on pg. 198—it gives an elementary version of the Morse theory proof of the Cartan-Hadamard Theorem (Theorem IV.8-13) in Milnor {2}.

Guggenheimer, H. W.

- {1} *Differential Geometry*, McGraw-Hill, New York, 1963 (MR 27 #6194).

The 1-parameter subgroups of the group of isometries of the hyperbolic plane are discussed in detail (pp. 276–278). The discussion of space curves in affine geometry (pp. 170–172) is especially interesting; it turns out that the analysis given in Chapter III.2 is not really adequate. Attention might also be called to exercise 12 on pg. 288.

Guillemin, V. and Pollack, A.

- {1} *Differential Topology*, Prentice Hall, Englewood Cliffs, N.J., 1974 (MR 50 #1276).

A beautiful extension, and complement, of Milnor {1}.

Haack, W.

- {1} *Elementare Differentialgeometrie*, Birkhäuser Verlag, Basel, 1955 (MR 18 596).

This seems to be the only book that treats infinitesimal bending by means of moving frames, and is the source for the material on pp. V.225–226. The author's definition seems to imply that the infinitesimal bend-

ing is non-trivial if and only if $\dot{\psi}_1^3$ or $\dot{\psi}_2^3$ is not identically zero, but I don't see offhand how to prove it. This fact is implicitly used in Problem V.12-4, which is a theorem of Weingarten (pg. 217); an analogous theorem (pg. 218) tells when "the lines of curvature remain lines of curvature in an infinitesimal bending" (i.e., when $\dot{l}_{12} = 0$). Eisenhart {2; 387} may be consulted for another proof, as well as for many examples of the "isothermal" surfaces arising in Problem V.12-4 (in particular, minimal surfaces are isothermal). For some modern theorems on infinitesimal bending that are basically translations of results about PDE's, see pp. 219–226. Finally, there is a proof (pp. 131–133) that only surfaces of constant curvature can be mapped conformally on the plane in such a way that the geodesics go to circles or straight lines. The proof is really unsatisfactory, for it uses certain lines that only exist on complex surfaces, and thus requires that the original surface be analytic. (I also know of no example to show that conformality of the mapping is a necessary hypothesis.) Bibliography of about 60 items.

Hadwiger, H.

{1} *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*, Springer-Verlag, Berlin, 1957 (MR 21 #1561).

Halperin, S.

see Greub, W.

Hartshorne, R.

{1} *Foundations of Projective Geometry*, W. A. Benjamin, New York, 1967 (MR 36 #5801).

Hawking, S. W. and Ellis, G. F. R.

{1} *The Large Scale Structure of Space-time*, Cambridge University Press, Cambridge, 1973 (MR 54 #12154).

Helgason S.

{1} *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962 (MR 26 #2986).

See Chapter I, §12 for a completely invariant definition of sectional curvature which does not involve the curvature tensor.

Hermann, R.

{1} *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968 (MR 38 #1635).

{2} *Geometry, Physics, and Systems*, Dekker, New York, 1973 (MR 58 #13104).

Hicks, N. J.

- {1} *Notes on Differential Geometry*, Van Nostrand, Princeton, N. J., 1965 (MR 31 #3936).

On pp. 122–123 there are some rigidity results of a simple nature. On pg. 154 there is a direct proof that the conjugate values of a geodesic are isolated, and pp. 168–169 classify the constant curvature simply-connected manifolds as is done in Bishop and Crittendon {1}.

Hilbert, D.

see Courant, R.

Hilbert, D. and Cohn-Vossen, S.

- {1} *Geometry and the Imagination*, Chelsea, New York, 1952 (MR 13 766).

An almost universally admired book, that discusses the visual and intuitive aspects of geometry rather than developing machinery and proofs. A refreshing view of the geometry of curves and surfaces in the differential geometry section, with references to all sorts of material that you won't find anywhere else.

Hlavatý, V.

- {1} *Differentialgeometrie der Kurven und Flächen und Tensorrechnung*, P. Noordhoff, Groningen, 1939.

- {2} *Differential Line Geometry*, P. Noordhoff, Groningen, 1953 (MR 15 252).

Hu, S. T.

- {1} *Differentiable Manifolds*, Holt, Rinehart and Winston, New York, 1969 (MR 39 #6343).

Huck, H., Roitzsch, R., Simon, U., Vortisch, W., Walden, R., Wegner, B., and Wendland, W.

- {1} *Beweismethoden der Differentialgeometrie im Grossen*, Springer-Verlag, Berlin, 1973 (MR 51 #6666).

Rigidity theorems using integral formulas and the index method.

Ishihara, S.

see Yano, K.

John, F.

- {1} *Partial Differential Equations*, Springer-Verlag, Berlin, 1971 (MR 46 #3960).

The source for most of the material in the beginning of Chapter V.10. Probably the best introduction to PDE's, with a good bibliography for more advanced study.

Jost, J.

- {1} *Riemannian Geometry and Geometric Analysis*, 2nd ed., Springer-Verlag, Berlin, 1998 (MR 99g:53025).

Kähler, E.

- {1} *Einführung in die Theorie der Systeme von Differentialgleichungen*, Chelsea, New York, 1949.

Kamber, F. and Tondeur, P.

- {1} *Flat Manifolds*, Springer-Verlag, Berlin, 1968 (MR 38 #6618).

Killing, W.

- {1} *Die Nicht-euklidischen Raumformen in Analytischer Behandlung*, G. B. Teubner, Leipzig, 1885.

Klingenberg, W.

- {1} *Riemannian geometry*, 2nd ed., Walter de Gruyter & Co, Berlin, 1995 (MR 95m:53003).
- {2} *Lectures on Closed Geodesics*, Springer-Verlag, New York, 1978 (MR 57 #17563).
- {3} *A Course in Differential Geometry*, Springer-Verlag, New York, 1978 (MR 57 #13702).
- {4} *Eine Vorlesung über Differentialgeometrie*, Springer-Verlag, Berlin, 1973 (MR 54 #3598).

see also Flaschel, P. and Gromoll, D.

Kobayashi, S.

- {1} *Transformation Groups in Differential Geometry*, Springer-Verlag, Berlin, 1972 (MR 50 #8360).
- {2} *Hyperbolic Manifolds and Holomorphic Mappings*, Dekker, New York, 1970 (MR 43 #3503).

Kobayashi, S. and Nomizu, K.

- {1} *Foundations of Differential Geometry*, 2 vols., Interscience, New York, 1963, 1969 (MR 27 #2945; 38 #6501).

This will probably become the standard reference for this generation. A complete treatment of the foundations, and the definitive exposition of the principal bundle point of view. Not exactly the sort of book to read like a novel, but one you should certainly have. There is a very large bibliography:

Kobayashi, S., Obata, M., and Takahashi, T. (eds.)

- {1} *Differential Geometry*, Kinokuniya, Tokyo, 1972 (MR 48 #2941).

Kodaira, K.

see Morrow, J.

Kolmogorov, A. N.

see Alexandrov, A. D.

Kreyszig, E.

- {1} *Introduction to Differential Geometry and Riemannian Geometry*, University of Toronto Press, Toronto, 1968 (MR 37 #2096).

See pg. 267 for the theorem on mapping of surfaces mentioned under Haack {1}. There are many interesting references to classical papers.

Kruppa, E.

- {1} *Analytische und Konstruktive Differentialgeometrie*, Springer, Wien, 1957 (MR 19 165).

Kulk, W. v. d.

see Schouten, J. A.

Lanczos, C.

- {1} *Space Through the Ages*, Academic Press, New York, 1970 (MR 42 #5747).

Historical, somewhat popularized, account.

Lane, E. P.

- {1} *A Treatise on Projective Differential Geometry*, University of Chicago Press, Chicago, 1942 (MR 4 114).

Large bibliography.

Lang, S.

- {1} *Differentiable Manifolds*, Addison-Wesley, Reading, Mass., 1972 (MR 55 #4241).

The standard reference, with everything done neatly and cleanly, for how things go in the infinite dimensional case.

- {2} *Differential and Riemannian Manifolds*, Springer-Verlag, New York, 1995 (MR 96d:53001).

Again, for the infinite dimensional case.

Lascoux, A. and Berger, M.

- {1} *Variétés Kähleriennes Compactes*, Springer-Verlag, Berlin, 1970 (MR 43 #3979).

Laugwitz, D.

- {1} *Differential and Riemannian Geometry*, Academic Press, New York, 1965 (MR 30 #2406).

Despite the usual problems with differentials, this book is a rather nice introduction to classical differential geometry, as well as to modern material. On pg. 131 there is an interesting formula for ΔN , where N is the

normal map of an immersion $f: M^2 \rightarrow \mathbb{R}^3$ (compare with the formula $\Delta f = 2H\nu$, pg. IV.136); note that the first term in Luagwitz' formula is just $-2\langle dH, df \rangle$. The references on this page may also be of interest. §15 treats not only Finsler metrics, but also "systems of paths" (vector fields on the tangent bundle).

- {2} *Differentialgeometrie in Vektorräumen*, Friedr. Vieweg & Sohn, Braunschweig, 1965 (MR 32 #406).

Lavrent'ev, M. A.

see Alexandrov, A. D.

Lawson, H. B. Jr.

- {1} *Lectures on Minimal Submanifolds*, Volume 1, Publish or Perish, Inc., 1979 (MR 82d:53035b).

A good review source for minimal surfaces and higher dimensional analogues, with a discussion of the Plateau problem, and bibliography.

- {2} *Minimal Varieties in Real and Complex Geometry*, University of Montréal Press, Montréal, 1972 (MR 57 #13798).

Leichtweiss, K.

see Blaschke, W.

Lelong-Ferrand, J.

- {1} *Géométrie Différentielle*, Masson, Paris, 1963 (MR 27 #648).

Levi-Civita, T.

- {1} *The Absolute Differential Calculus*, Blackie & Son, London, 1961.

Of historical interest only.

Lichnerowicz, A.

- {1} *Théorie Globale des Connexions et des Groupes d'Holonomie*, Edizioni cremonese, Rome, 1955 (MR 19 453).
- {2} *Géométrie des Groupes de Transformations*, Dunod, Paris, 1958 (MR 23 #A1329).

Lie, S. and Scheffers, G. W.

- {1} *Vorlesungen über Continuierliche Gruppen*, B. G. Teubner, Leipzig, 1893.
- {2} *Geometrie der Berührungstransformationen*, B. G. Teubner, Leipzig, 1896.

Loos, O.

- {1} *Symmetric Spaces*, 2 vols., W. A. Benjamin, New York, 1969 (MR 39 #365a, b).

Lovelock, D.

see Rund, H.

Lyusternik, L. A.

- {1} *Shortest Paths. Variational Problems*, Pergamon Press, Macmillan, New York, 1964 (MR 31 #2644).

A beautiful little book that gives elementary treatments of variational problems. There is a geometric proof of Clairaut's theorem in which the surface of revolution is approximated by a union of frustra of cones, for which the theorem is checked directly.

MacKenzie, R. E.

see Auslander, L.

Malliavin, P.

- {1} *Géométrie Différentielle Intrinsèque*, Hermann, Paris, 1972 (MR 57 #13704).

Marsden, J.

see Abraham, R.

Matsumoto, M.

- {1} *The Theory of Finsler Connections*, Okayama University, Okayama, 1970 (MR 42 #2409).

Matsushima, Y.

- {1} *Differentiable Manifolds*, Marcel Dekker, New York, 1972 (MR 49 #11553).

Mazet, E.

see Berger, M.

McCleary, J.

- {1} *Geometry from a Differentiable Viewpoint*, Cambridge University Press, Cambridge, 1994 (MR 95m:53001).

Menger, K.

see Blumenthal, L. M.

Meyer, W.

see Gromoll, D.

Millman, R. S. and Parker, G. D.

- {1} *Elements of Differential Geometry*, Prentice Hall, Englewood Cliffs, N.J., 1977 (MR 56 #1208).

Milnor, J. W.

- {1} *Topology from the Differentiable Viewpoint*, University Press of Virginia, Charlottesville, Virginia, 1965 (MR 37 #2239).

This book is not really about differential geometry at all, but about differential topology. But this is a field of related interest, and besides, anything written by Milnor is beautiful. See, in particular, §2 for the hard version of Sard's theorem.

- {2} *Morse Theory*, Princeton University Press, Princeton, N.J., 1963 (MR 31 #6249).

Although this book is really devoted to more advanced material from differential topology, a substantial part contains material from differential geometry. See, in particular, §6 and §19, where the Cartan-Hadamard theorem is proved using Morse theory (compare the remarks under Gromoll, Klingenberg, and Meyer {1}).

Milnor, J. W. and Stasheff, J. D.

- {1} *Characteristic Classes*, Princeton University Press, Princeton, N.J., 1974 (MR 55 #13428).

The first characteristic classes considered here are the Stiefel-Whitney classes, with \mathbb{Z}_2 coefficients. Since the actual construction of these classes, in §8, involves the Steenrod squares, it might be advisable simply to skip this section. §§9–15 proceed to define the Euler, Pontryagin, and Chern classes in the way mentioned in the “Valedictory” to Chapter V.13. The remaining §§ give applications beyond those interspersed in the preceding ones. We should explicitly mention that §11 identifies the Euler class of TM without invoking triangulations. §6 describes the cell structure for the Grassmannians by means of Schubert varieties. This cell structure was originally used by Ehresmann [1], [2] to compute the (integral) cohomology of the Grassmannians. From this one immediately obtains the real cohomology, which of course agrees with our calculations in Chapter V.13. Naturally it would be of interest to connect the two calculations directly, by determining the integrals over the Schubert varieties of the forms giving the Chern, Pontryagin, and Euler classes. I know of no reference to such calculations, except for the brief note of Chern [4] and the mimeographed notes of Chern {2}.

Misner, C. W., Thorne, K. S., and Wheeler, J. A.

- {1} *Gravitation*, Freeman, San Francisco, 1973 (MR 54 #6869).

Morgan, F.

- {1} *Riemannian Geometry. A beginner's guide*, 2nd ed., A K Peters, Ltd., Wellesley, Mass., 1998 (MR 98i:53001).

Mainly an intuitive introduction to Riemannian geometry, with some additional esoteric material.

Morrow, J. and Kodaira, K.

- {1} *Complex Manifolds*, Holt, Rinehart and Winston, New York, 1971 (MR 46 #2080).

Munkres, M. R.

- {1} *Elementary Differential Topology*, Princeton University Press, Princeton, N.J., 1963 (MR 29 #623).

This little book proves almost all the facts about differentiable manifolds which arise in the description of C^k -manifolds, and which a differential geometer might want to know, even though they are not part of differential geometry *per se*.

Narasimhan, R.

- {1} *Analysis on Real and Complex Manifolds*, Masson, Paris, 1968 (MR 40 #4972).

Nelson, E.

- {1} *Tensor Analysis*, Princeton University Press, Princeton, N.J., 1967.

An unorthodox approach to tensors, which has some neat things. See, in particular, the discussion of the bracket (pp. 30–36), the interpretation of curvature (pg. 77), and the approach to the operator δ which uses Proposition IV.7-62 as the definition, instead of invoking the $*$ operator (pp. 96–100).

Nomizu, K. and Sasaki, T.

- {1} *Affine Differential Geometry*, Cambridge University Press, Cambridge, 1994 (MR 96e:53014).

see also Kobayashi, S.

Obata, M.

see Kobayashi, S.

O'Neill, B.

- {1} *Elementary Differential Geometry*, Academic Press, New York, 1966 (MR 34 #3444).

This book expounds moving frames for undergraduates. Mention might be made of the material on pp. 330–333 and exercises 10–13 on pp. 337–338, generalizing our discussion of geodesics on surfaces of revolution (pp. III.212–216). Also exercise 8 on pg. 387 gives the Gauss-Bonnet theorem for surfaces-with-boundary.

- {2} *Semi-Riemannian Geometry*, Academic Press, Inc., New York, 1983 (MR 85f:53002).

Osserman, R.

- {1} *Survey of Minimal Surfaces*, Van Nostrand Reinhold, New York, 1969 (MR 41 #934).

A good place to learn more about minimal surfaces, and minimal submanifolds, with a large bibliography. I would probably have gone out of my mind trying to write Chapter IV.9 if it hadn't been for this book, which clears up things that standard texts have contentedly left in a muddle for years. For example, nowhere else are the poles of g considered in the Enneper-Weierstrass representation of minimal surfaces. Also, almost all classical books consider all minimal surfaces to be complex surfaces. This is valid, since minimal surfaces are analytic, and can therefore be complexified, but it usually leads to results that have no real geometric significance. For example, a theorem of Lie (cf. Blaschke {1; v.1, pg. 236}, Haack {1; pg. 140}, or Kreyszig {1; pg. 244}) says that every minimal surface is a “translation surface”, of the form $(s, t) \mapsto f(s) + g(t)$. But this is only true if f and g are complex-valued; it is easy to see that the only minimal surfaces of this form for \mathbb{R}^3 -valued functions f and g is Scherk's minimal surface (in fact, that's how the surface was discovered).

Parker, G. D.

see Millman, R. S.

Pars, L.

- {1} *A Treatise on Analytic Dynamics*, Wiley, New York, 1965.

Petersen, P.

- {1} *Riemannian Geometry*, Springer-Verlag, New York, 1998 (MR 98m:53001).

Petrov, A. Z.

- {1} *Einstein Spaces*, Pergamon Press, Oxford, 1969 (MR 39 #6225).

Pham Mau Quan

- {1} *Introduction à la Géométrie des Variétés Différentiables*, Dunod, Paris, 1969 (MR 39 #3415).

Pogorelov, A. V.

- {1} *Die Eindeutige Bestimmung Allgemeiner Konvexer Flächen*, Akademie-Verlag, Berlin, 1956 (MR 18 330).
- {2} *Die Verbiegung Konvexer Flächen*, Akademie-Verlag, Berlin, 1957 (MR 19 305).

{3} *Topics in the Theory of Surfaces in Elliptic Space*, Gordon and Breach, New York, 1961 (MR 26 #6908).

{4} *Einige Untersuchungen zur Riemannschen Geometrie im Grossen*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1960 (MR 22 #5946).

Concerning the imbedding of 2-dimensional Riemannian manifolds in a given 3-dimensional Riemannian manifold.

{5} *Extrinsic Geometry of Convex Surfaces*, American Mathematical Society, Providence, R.I., 1973 (MR 49 #11439; 39 #6222).

{6} *Monge-Ampère Equations of Elliptic Type*, P. Noordhoff, Groningen, 1964 (MR 31 #4993).

{7} *Differential Geometry*, P. Noordhoff, Groningen, 1959 (MR 22 #4990).

{8} *The Minkowski Multidimensional Problem*, V. H. Winston & Sons, Washington, D.C., John Wiley & Sons, New York, 1978 (MR 57 #17572).

Pollock, A.

see Guillemin, V.

Poston, T.

see Dodson, C. T. J.

Rado, T.

{1} *On the Problem of Plateau*, Springer, Berlin, 1933.

Raschewski, P. K.

{1} *Riemannsche Geometrie und Tensoranalysis*, Deutscher Verlag der Wissenschaften, Berlin, 1959 (MR 21 #2258).

Reichardt, H.

see Blaschke, W.

Rham, G. de

{1} *Variétés Différentiables. Formes, courants, formes harmoniques*, Hermann, Paris, 1955 (MR 16 957).

Contains, among other things, a proof of Hodge's theorem by means of "currents"—these are essentially differential forms whose coefficients are distributions (in the analysts' sense).

Also available in English translation:

Differentiable Manifolds, Springer-Verlag, Berlin-New York, 1984 (MR 85m:58005).

Riemann, B.

{1} *Collected Works*, Dover, New York, 1953 (MR 14 610).

Rinow, W.

- {1} *Die Innere Geometrie der Metrischen Räume*, Springer, Berlin, 1961 (MR 23 #A1290).

Roitzsch, R

see Huck, H.

Rund, H.

- {1} *The Differential Geometry of Finsler Spaces*, Springer, Berlin, 1959 (MR 21 #4462).
- {2} *The Hamilton-Jacobi Theory in the Calculus of Variations; Its role in mathematics and physics*, Van Nostrand, London, 1966 (MR 37 #5752).

Rund, H. and Lovelock, D.

- {1} *Tensors, Differential Forms, and Variational Principles*, John Wiley & Sons, New York, 1975 (MR 57 #13703).

Rund, H. and Forbes, W. F. (eds.)

- {1} *Topics in Differential Geometry*, Academic Press, New York, 1976 (MR 53 #1430).

Ruse, H. S., Walker, A. G., and Willmore, T. J.

- {1} *Harmonic Spaces*, Edizioni Cremonese, Rome, 1961 (MR 25 #5456).

Sachs, R. K. and Wu, H.-H.

- {1} *General Relativity for Mathematicians*, Springer-Verlag, New York, 1977 (MR 58 #20239a).

Salkowski, E.

- {1} *Affine Differentialgeometrie*, de Gruyter, Berlin, 1934.

Santalo, L. A.

- {1} *Introduction to Integral Geometry*, Hermann, Paris, 1953 (MR 15 736).

Sasaki, T.

see Nomizu, K.

Sauer, R.

- {1} *Differenzengeometrie* [sic], Springer, Berlin, 1970 (MR 41 #7544).

Scheffers, G. W.

- {1} *Anwendung der Differential- und Integral-Rechnung auf Geometrie*, 2 vols., Veit & Co., Leipzig, 1901–1902.

A classical book with quite extended discussions of various topics, too numerous to be listed here. Special mention may be made of the analytic determination of all geodesic maps between open subsets of the plane

(v.2, pp. 429–432). The unique features of the book have been mentioned under topic III.(g).

see also Lie, S.

Schild, A.

see Synge, J. L.

Schirokow, P. A. and Schirokow, A. P.

{1} *Affine Differentialgeometrie*, Teubner, Leipzig, 1962 (MR 27 #660).

Contains an extensive bibliography, and historical remarks in the forward.

Schouten, J. A.

{1} *Ricci-Calculus*, 2nd ed., Springer-Verlag, Berlin, 1954 (MR 16 521).

This book was written to show the great superiority of the classical notation. Reading it, one can see why differential geometry was once given up for dead. There are super-, sub-, pre- and post-scripts, including dots, brackets, etc. It is certainly a triumph of the printer's art, but is also supposed to be important for its considerations of non-symmetric connections (sometimes considered in general relativity). There is an enormous bibliography. Amusing footnotes on pp. 118, 160, 172.

Schouten, J. A. and Kulk, W. v. d.

{1} *Pfaff's Problems and its Generalization*, Clarendon Press, Oxford, 1949 (MR 11 179).

Schouten, J. A. and Struik, D. J.

{1} *Einführung in die Neueren Methoden der Differentialgeometrie*, 2nd ed., 2 vols., P. Noordhoff, N.V., Groningen-Batavia, 1935, 1938.

In volume 2 see pg. 78 for references to generalizations of the Beltrami-Enneper theorem (another reference is Hayden [1]), pp. 94–116 for a detailed classification of points in a submanifold, and pg. 136 for results related to the invariance of K_r when r is even (pg. IV.70). The notation, though classical, is quite manageable.

Schwartz, J. T.

{1} *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969 (MR 55 #6457).

{2} *Differential Geometry and Topology*, Gordon and Breach, New York, 1968.

See pg. 87 for a direct verification of the fact that all partial derivatives of g_{ij} at the center of a normal coordinate system are expressible in terms of the covariant derivatives of the curvature tensor (compare pp. IV.156–157). The original proof of this fact was by Vermeil [1]. By

the way, “normal coordinates” have been used for various tensors other than a Riemannian metric (cf. references at the end of Veblen {1; ch.6}, and also Weitzenböck {1; §§19–22}).

Segre, B.

- {1} *Some Properties of Differentiable Varieties and Transformations; with special reference to the analytic and algebraic Cases*, Springer, Berlin, 1957 (MR 16 679).

Sharpe, R. W.

- {1} *Differential Geometry. Cartan's generalization of Klein's Erlangen program*, Springer-Verlag, New York, 1997 (MR 98m:53033).

Simon, U.

see Huck, H.

Singer, I. M. and Thorpe, J. A.

- {1} *Lecture Notes on Elementary Topology and Geometry*, Scott, Foresman, Glenview, Ill., 1967 (MR 35 #4834).

Another non-classical introduction to surface theory for undergraduates, this time in terms of principal bundles. The trick is that for surface theory the principal bundle degenerates to the sphere bundle and everything is much easier. Nevertheless, the book is not easy going. Chapter 6 contains a weird proof of the de Rham theorem.

Śleboziński, W.

- {1} *Exterior Forms and Their Application*, Państwowe Wydawnictwo Naukowe, Warsaw, 1970 (MR 42 #672).

Sommerville, D. M. Y.

- {1} *Bibliography of Non-euclidean Geometry including the theory of parallels, the foundations of geometry, and space of n dimensions*, Harrison & Sons, London, 1911.

Sorani, G.

- {1} *An Introduction to Real and Complex Manifolds*, Gordon and Breach, New York, 1969 (MR 41 #6220).

Stasheff, J. D.

see Milnor, J. W.

Stehle, P.

see Corbin, H.

Sternberg, S.

- {1} *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, N.J., 1964 (MR 33 #1797).

The author's heart was really in the last chapter, on G -structures, and perhaps in the fourth chapter on the calculus of variations. But see also

pg. 45 for the hard version of Sard's theorem, pg. 63 for the Whitney imbedding theorem, and pg. 256 for Whitney's theorem on smooth homotopy of plane curves.

Stoker, J. J.

{1} *Differential Geometry*, Wiley-Interscience, New York, 1969 (MR 39 #2072).

Strubecker, K.

{1} *Differentialgeometrie*, 3 vols., Walter de Gruyter & Co., Berlin, 1964, 1969, 1969 (MR 16 954; 20 #4273; 21 #878).

Three pretty little paperback volumes with lots of details about elementary classical topics, and dozens of excellent pictures.

Struik, D. J.

{1} *Lectures on Classical Differential Geometry*, Addison-Wesley, Reading, Mass., 1961 (MR 12 227).

See pp. 153–156 for a simple classical proof of the Gauss-Bonnet theorem.

see also Schouten, J. A.

Sulanke, R. and Wintgen, P.

{1} *Differentialgeometrie und Faserbündel*, Birkhäuser Verlag, Basel, 1972 (MR 54 #1274).

Švec, A.

{1} *Projective Differential Geometry of Line Congruences*, Czechoslovak Academy of Sciences, Prague, 1965 (MR 33 #7949).

Synge, J. L. and Schild, A.

{1} *Tensor Calculus*, University of Toronto Press, Toronto, 1962 (MR 11 400).

Takahashi, T.

see Kobayashi, S.

Thomas, T. Y.

{1} *The Differential Invariants of Generalized Space*, Cambridge University Press, Cambridge, 1937.

{2} *Concepts from Tensor Analysis and Differential Geometry*, 2nd ed., Academic Press, New York, 1965 (MR 32 #4623).

Thorne, K. S.

see Misner, C. W.

Thorpe, J. A.

see Singer, I. M.

Tondeur, P.

see Kamber, F.

Vaisman, I.

{1} *Cohomology and Differential Forms*, Dekker, New York, 1973 (MR 81i:53003).

The sheaf theory here is more advanced than in Warner {1}. There is a brief account of the theory of Allendoerfer and Eels [1], which describes the *integral* cohomology of M in terms of forms with singularities.

Vanstone, R.

see Greub, W.

Veblen, O.

{1} *Invariants of Quadratic Differential Forms*, Cambridge University Press, Cambridge, 1927.

Veblen, O. and Whitehead, J. H. C.

{1} *The Foundations of Differential Geometry*, Cambridge University Press, Cambridge, 1932.

Vortisch, W.

see Huck, H.

Vranceanu, G.

{1} *Lectii de Geometrie Diferentiala*, 4 vols., Editura Academiei Republicii Socialiste România, Bucharest, 1968 (MR 39 #6181).

The first volume was originally written in French, and then translated into Rumanian, the language in which the last three volumes were originally written. The second and third volumes have been translated into French, and the first two volumes have been translated into German. Various changes have been made in some of the translations. The review cited above is for Volume 4 only, but it contains a complete list of all other reviews of previous volumes.

Walden, R.

see Huck, H.

Walker, A. G.

see Ruse, H. S.

Warner, F. W.

{1} *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Co., Glenview, Ill., 1971 (MR 45 #4312).

The first part of this book treats manifold theory and Lie groups concisely, but with all the necessary details. The relationship between the functors $\mathcal{T}^k(V)$ and $\Omega^k(V)$ and the algebraists' $\bigotimes^k V$ and $\Lambda^k V$ are spelled out in detail in Chapter 2. The last theorem of that chapter formally states a fact which we have frequently used (e.g., in the proof of

Theorem IV.7-20). The third chapter gives a little more detail on Lie groups. The adjoint representation is treated explicitly, as is the universal covering group of a Lie group. See also exercise 15 on pg. 134 for a proof that the exponential map of $GL(n, \mathbb{C})$ is onto and exercise 18 on pg. 135 for the classification of Abelian Lie groups.

The second and third parts of the book give connected presentations of material available nowhere else. The sheaf-theoretic proof of de Rham's theorem is given in the second part: Alexander-Spanier, Čech, and singular cohomologies are all mentioned, the isomorphism between de Rham and singular cohomology is shown to be given by integration, and the wedge product is shown to correspond to cup product.* (Although a simple proof of the de Rham theorem was outlined in Problem I.11-14, the sheaf-theoretic proof explains, in some sense, "why" the theorem is true, for it shows that only certain basic facts ($d^2 = 0$, the local converse, and existence of partitions of unity) matter, while the details of how they are obtained is irrelevant.) The third part of the book gives a completely self-contained elementary treatment of the Hodge theorem.

Wasserman, R.

{1} *Tensors and Manifolds*, The Clarendon Press, Oxford University Press, New York, 1992 (MR 93h:53002).

The reader might be concerned, as I was, whether this holds for our wedge product, with the factor $(p+q)!/p!q!$, or for the wedge product without this factor. The answer is that it holds for either! One way to see this is the following. In $\Lambda^k(V)$, regarded as a quotient vector space of $\bigotimes^k V$, there is a natural \wedge product ($v_1 \wedge \cdots \wedge v_k$ is the residue class of $v_1 \otimes \cdots \otimes v_k$), and this wedge product on each $\Lambda^k(M_p^)$ gives a wedge product on differential forms if we regard a k -form as a function $p \mapsto \omega(p) \in \Lambda^k(M_p^*)$. It is then this wedge product that corresponds to cup product under the de Rham isomorphism, so long as we define integration so that the integral of $f dx^1 \wedge \cdots \wedge dx^n$ on \mathbb{R}^n is just the ordinary integral of f . Now we can also get a wedge product on $\Omega^k(V)$ by means of an isomorphism of $\Omega^k(V)$ with $\Lambda^k(V^*)$. One isomorphism will give our wedge product, while another will give the product without the factor $(p+q)!/p!q!$, but in either case the wedge product will correspond to cup product. If this explanation seems too paradoxical, the following may help. For a k -form ω on M and a singular k -cube $c: [0, 1]^k \rightarrow M$, we define $\int_c \omega$ as the ordinary integral of f over $[0, 1]^k$ where

$$c^*\omega = f dx^1 \wedge \cdots \wedge dx^k.$$

In this definition, we naturally use the same \wedge in $dx^1 \wedge \cdots \wedge dx^k$ as we use in taking the product $\omega \wedge \eta$ on our manifold M ; if we change the definition of \wedge , then we also end up changing the definition of $\int_c \omega$, and the change involves just the right factor so that \wedge still corresponds to cup product.

Watson, G. H.

see Whittaker, E. T.

Weatherburn, C. E.

- {1} *An Introduction to Riemannian Geometry and the Tensor Calculus*, Cambridge University Press, Cambridge, 1957.

The historical note at the end may be of interest—it contains some names that you have probably never even seen before.

Weber, W. C. and Goldberg, S. I.

- {1} *Conformal Deformations of Riemannian Manifolds*, Queen's University, Kingston, Ontario, 1969 (MR 40 #1938).

Wegner, B.

see Huck, H.

Weil, A.

- {1} *Introduction à l'Étude des Variétés Kähleriennes*, Hermann, Paris, 1958 (MR 22 #1921).

Weinberg, S.

- {1} *Gravitation and Cosmology*, Wiley, New York, 1972.

Weiss, G. L.

see Boothby, W.

Weitzenböck, R.

- {1} *Invariantentheorie*, P. Noordhoff, Groningen, 1923.

Famous for the message contained in the initial letters of the sentences in the forward. In addition to the material of a differential geometric nature, some of the classical invariant theory might be of interest, for example, apolarity conditions. One might also consult Dieudonné and Carrell {1} (or [1]).

Wells, R. O. Jr.

- {1} *Differential Analysis on Complex Manifolds*, Prentice-Hall, Englewood Cliffs, N.J., 1973 (MR 58 #24309a).

Wendland, W.

see Huck, H.

Weyl, H.

- {1} *The Classical Groups, their invariants and representations*, 2nd ed., Princeton University Press, Princeton, N.J., 1953 (MR 1 42).

This book certainly qualifies as a classic, if unreadability is one of the criteria. It was the source for the proof of the first main theorem of invari-

ance theory for $O(n)$ and $SL(n, k)$. One might also consult Dieudonné and Carrell {1} (or [1]).

Wheeler, J. A.

see Misner, C. W.

Whitehead, J. H. C.

see Veblen, O.

Whittaker, E. T.

{1} *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., Cambridge University Press, Cambridge, 1959 (MR 21 #2381).

Whittaker, E. T. and Watson, G. N.

{1} *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, 1962 (MR 31 #2375).

Wilczynski, E. J.

{1} *Projective Differential Geometry of Curves and Ruled Surfaces*, B. G. Teubner, Leipzig, 1906.

Willmore, T. J.

{1} *An Introduction to Differential Geometry*, Oxford University Press, London, 1959 (MR 28 #2482).

{2} *Total Curvature in Riemannian Geometry*, Wiley, New York, 1982 (MR 84f:53034).

{3} *Riemannian Geometry*, The Clarendon Press, Oxford University Press, New York, 1993 (MR 95e:53002).

see also Ruse, H. S.

Wintgen, P.

see Sulanke, R.

Wolf, J. A.

{1} *Spaces of Constant Curvature*, 5th ed. Publish or Perish, Houston, Texas, 1984 (MR 88k:53002; 36 #829).

After a rapid run through Riemannian geometry, this book gets down to the main task of classifying complete manifolds of constant curvature. This is basically an algebraic problem, since all such manifolds have the standard simply-connected examples as their universal covering spaces. The case of n -dimensional manifolds M of constant positive curvature is easy for n even: M is S^n or \mathbb{P}^n (pg. 74 or Kobayashi and Nomizu {1; v.1, Note 4}). Other cases are not so easy, and all sorts of interesting material gets woven together in the search for the final classification.

Wright, J. E.

- {1} *Invariants of Quadratic Differential Forms*, Cambridge University Press, Cambridge, 1908.

Wu, H.-H.

- {1} *The Equidistribution Theory of Holomorphic Curves*, Princeton University Press, Princeton, N.J., 1970 (MR 42 #7951).

see also Sachs, R. K.

Yaglom, I. M. and Boltyanskiĭ, V. G.

- {1} *Convex Figures*, Holt, Rinehart and Winston, New York, 1961 (MR 23 #A1283).

Yano, K.

- {1} *Integral Formulas in Riemannian Geometry*, Dekker, New York, 1970 (MR 44 #2174).

Large bibliography.

- {2} *Differential Geometry on Complex and Almost Complex Spaces*, Macmillan, New York, 1965 (MR 32 #4635).

- {3} *The Theory of Lie Derivatives and its Applications*, North-Holland, New York, 1957 (MR 19 576).

Large bibliography.

Yano, K. and Bochner, S.

- {1} *Curvature and Betti Numbers*, Princeton University Press, Princeton, N.J., 1953 (MR 15 989).

Yano, K. and Ishihara, S.

- {1} *Tangent and Cotangent Bundles: Differential Geometry*, Dekker, New York, 1973 (MR 50 #3142).

Zalgaller, V. A.

see Alexandrov, A. D.

C. JOURNAL ARTICLES

Abresch, U.

- [1] *Constant mean curvature tori in terms of elliptic functions*, J. Reine Angew. Math. **374** (1987), 169–192 (MR 88e:53006).

Alexandrov, A. D.

- [1] *Uniqueness theorems for surfaces in the large. I*, Vestnik Leningrad. Univ. **11** (1956), no. 19, 5–17 [Russian] (MR 19 167); Amer. Math. Soc. Transl. (2) **21** (1962), 341–354 (MR 27 #698a).
- [2] *A characteristic property of spheres*, Ann. Mat. Pura Appl. (4) **58** (1962), 303–315 (MR 26 #722).
- [3] *Zur Theorie der gemischten Volumina von konvexen Körpern*, Mat. Sb. 2 **44** (1937), 947–972, 1205–1238; 3 **45** (1938), 27–46, 227–251 [Russian. German summary].
- [4] *On a class of closed surfaces*, Mat. Sb. **4** (1938), 69–77 [Russian].

Allendoerfer, C. B.

- [1] *The Euler number of a Riemannian manifold*, Amer. J. Math. **62** (1940), 243–248 (MR 2 20).
- [2] *Rigidity for spaces of class greater than one*, Amer. J. Math. **61** (1939), 633–644 (MR 1 28).
- [3] *The imbedding of Riemann spaces in the large*, Duke Math. J. **3** (1937), 317–333.

This paper is the source for Chapter IV.7, Addendum 4.

Allendoerfer, C. B. and Eels, J. Jr.

- [1] *On the cohomology of smooth manifolds*, Comment. Math. Helv. **32** (1958), 165–179 (MR 21 #868).

Allendoerfer, C. B. and Weil, A.

- [1] *The Gauss-Bonnet theorem for Riemannian polyhedra*, Trans. Amer. Math. Soc. **53** (1943), 101–129 (MR 4 169).

Barbosa, J. L. and do Carmo, M.

- [1] *On the size of a stable minimal surface in R^3* , Amer. J. Math. **98** (1976), no. 2, 515–528 (MR 54 #1292).

Beez, R.

- [1] *Zur Theorie des Krümmungsmasses von Mannigfaltigkeiten höhere Ordnung*, Zeitschrift für Mathematik und Physik **21** (1876), 373–401.

Bianchi, L.

- [1] *Sulle superficie a curvatura nulla in geometria ellittica*, Ann. Mat. Pura Appl. **24** (1896), 93–129.

Bieberbach, L.

- [1] *Hilberts Satz über Flächen konstanter negative Krümmung*, Acta Math. **48** (1926), 319–327.

This paper gives a rigorous proof of Hilbert's theorem which is closer to the original argument of Hilbert [1] than the one we have given (it also contains an unfounded criticism of Holmgren's proof).

Blaschke, W.

- [1] *Kreis und Kugel*, Jber. Deutsch. Math.-Verein. **24** (1915–1916), 195–207.

Bol, G.

- [1] *Über Nabelpunkte auf einer Eifläche*, Math. Z. **49** (1944), 389–410 (MR 7 29).

Carmo, M. do and Lima, E.

- [1] *Isometric immersions with semi-definite second quadratic forms*, Arch. Math. (Basel) **20** (1969), 173–175 (MR 39 #6214).
 [2] *Immersions of manifolds with non-negative sectional curvatures*, Bol. Soc. Brasil. Mat. **2** (1972), 9–22 (MR 48 #7170).

Carmo, M. do and Warner, F. W.

- [1] *Rigidity and convexity of hypersurfaces in spheres*, J. Differential Geometry **4** (1970), 133–144 (MR 42 #1014).

Note that the assertion, on pg. 140, that \tilde{x} and \tilde{y} are imbeddings needs some argument. (See the last part of the statement of Proposition V.12–26; as far as I can tell, Theorem 1 of Chapter IV of Pogorelov {3} does not claim that \tilde{x} and \tilde{y} are always immersions, as asserted in the editor's footnote, but only that their images have a tangent space at each point.)

see also Barbosa, J. L.

Carrell, J. B.

see Dieudonné, J. A.

Cartan, É.

- [1] *La déformation des hypersurfaces dans l'espace euclidien réel à n dimensions*, Bull. Soc. Math. France **44** (1916), 65–99.
 [2] *Les surfaces qui admettent une seconde forme fondamentale donnée*, Bull. Sci. Math. (2) **67** (1943), 8–32 (MR 7 30).
 [3] *Sur les variétés de courbure constante d'un espace euclidien ou non euclidien*, Bull. Soc. Math. France **47** (1919), 125–160; **48** (1920), 132–208.

Chern, S.-S.

- [1] *Integral formulas for hypersurfaces in Euclidean space and their applications to uniqueness theorems*, J. Math. Mech. **8** (1959), 947–956 (MR 22 #4997).
- [2] *A simple intrinsic proof of the Gauss-Bonnet formula for closed Riemannian manifolds*, Ann. of Math. (2) **45** (1944), 747–752 (MR 6 106).

See also the next paper, especially for the formula for manifolds-with-boundary.

- [3] *On the curvatura integra in a Riemannian manifold*, Ann. of Math. (2) **46** (1945), 674–684 (MR 7 328).
- [4] *On the characteristic classes of Riemannian manifolds*, Proc. Nat. Acad. Sci. U.S.A. **33** (1947), 78–82 (MR 8 490).
- [5] *On a theorem of algebra and its geometrical application*, J. Indian Math. Soc. **8** (1944), 29–36 (MR 6 216).
- [6] *An elementary proof of the existence of isothermal parameters on a surface*, Proc. Amer. Math. Soc. **6** (1955), 771–782 (MR 16 856).

This is the standard reference, but for more details I used Bers' NYU notes on Riemann surfaces (1957–1958). See also the papers cited in the review, and in the immediately preceding review on the same page.

Chern, S.-S. and Lashof, R. K.

- [1] *On the total curvature of immersed manifolds*, Amer. J. Math. **79** (1957), 306–318 (MR 18 927); *II*, Michigan Math. J. **5** (1958), 5–12 (MR 20 #4301).

Christoffel, E. B.

- [1] *Ueber die Transformation der homogenen Differentialausdrücke zweiten Grades*, J. Reine Angew. Math. **70** (1869), 46–70.

I could not resist mentioning this extraordinarily impressive paper, which appeared just one year after the publication of Riemann's inaugural lecture. Christoffel defines covariant differentiation, uses it to define the curvature tensor, and solves the problem of determining when two Riemannian manifolds are locally isometric. This is done essentially as in chapter V.7, Addendum 3, although the argument is not phrased in terms of the bundle of frames—the latter formulation comes from Cartan {3}.

Cohn-Vossen, S.

- [1] *Unstarre geschlossene Flächen*, Math. Ann. **102** (1929), 10–29.
- [2] *Zwei Sätze über die Starrheit der Eiflächen*, Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl. 1927, 125–134.

For another proof of the infinitesimal bendability of convex surfaces with a disc deleted, see Süss [1].

Courant, R.

- [1] *Soap film experiments with minimal surfaces*, Amer. Math. Monthly **47** (1940), 167–174 (MR 1 270).

Courant, R. and Lax, P.

- [1] *On nonlinear partial differential equations with two independent variables*, Comm. Pure Appl. Math. **2** (1949), 255–273 (MR 11 441).

Note that the region R_δ on pg. 263 is defined incorrectly—it has very steep sides, rather than the opposite. In the footnote on pg. 260 the authors make a more serious error; it amounts to asserting that the matrix A on pp. V.75–76 need not be assumed invertible, because this can always be achieved by an affine transformation of the plane. But, in fact, this works only if the dimension of $\ker A$ is only 1. Finally, the process by which the authors reach their equation (4.6) on pg. 261 is to me mysteriously complicated. For the arguments on pp. V.77–78, I used Lax's NYU notes on PDE's (1949–1950).

Cutler, E. H.

- [1] *On the curvatures of a curve in Riemann space*, Trans. Amer. Math. Soc. **33** (1931), 832–838.

This paper contains a generalization of Theorem IV.7-9: Let N be an arbitrary Riemannian manifold, let $M \subset N$ be a j -dimensional totally geodesic submanifold, and let $c: [a, b] \rightarrow N$ be an arclength parameterized curve with $\kappa_1, \dots, \kappa_{j-1}$ nowhere zero, and κ_j everywhere zero. Suppose that $c(a) \in M$ and $\mathbf{v}_i(a) \in M_{c(a)}$ for $i = 1, \dots, j$. Then $c([a, b]) \subset M$.

Dieudonné, J. A. and Carrell, J. B.

- [1] *Invariant theory, old and new*, Advance in Math. **4** (1970), 1–80 (MR 41 #186).

Also available in book form, see Dieudonné, J. A. {1}.

do Carmo, M.

see Carmo, M. do

Dolbeault-Lemoine, S.

- [1] *Sur la déformabilité des variétés plangées dans espace de Riemann*, Ann. Sci. Ecole Norm. Sup. (3) **73** (1956), 357–438 (MR 18 819).

Eels, J. Jr.

see Allendoerfer, C. B.

Efimov, N. V.

- [1] *Generation of singularities on surfaces of negative curvature*, Mat. Sb. **64** (106) (1964), 286–320 [Russian] (MR 29 #5203).

Ehresmann, C.

- [1] *Sur la topologie de certains espaces homogènes*, Annals of Math. **35** (1934), 396–443.
- [2] *Sur la topologie de certaines variétés algébriques réelles*, J. Math. Pures Appl. (9) **16** (1937), 69–100.

Fenchel, W.

- [1] *Über Krümmung und Windung geschlossener Raumkurven*, Math. Ann. **101** (1929), 238–252.
- [2] *On total curvatures of Riemannian manifolds: I*, J. London Math. Soc. **15** (1940), 15–22 (MR 2 20).

Fenchel, W. and Jessen, B.

- [1] *Mengenfunktionen und konvexe Körper*, Danske Videns. Selskab. Math-Fysiske Medd. **16** (1938), 1–31.

Fialkow, A.

- [1] *Hypersurfaces of a space of constant curvature*, Ann. of Math. **39** (1938), 762–785.

Firey, W. J.

- [1] *The determination of convex bodies from their mean radius of curvature functions*, Mathematika **14** (1967), 1–13 (MR 36 #788).

The next paper treats Christoffel's problem for non-differentiable convex surfaces.

- [2] *Christoffel's problem for general convex bodies*, Mathematika **15** (1968), 7–21 (MR 37 #5822).
- [3] *Intermediate Christoffel-Minkowski problems for figures of revolution*, Israel J. Math. **8** (1970), 384–390 (MR 42 #6719).

This paper considers the functions P_i for $1 < i < n$.

Flanders, H.

- [1] *Local theory of affine hypersurfaces*, J. Analyse Math. **15** (1965), 353–387 (MR 32 #403).

This paper uses moving frames to describe the (local) theory of hypersurfaces M^n of \mathbb{R}^{n+1} in special affine geometry. Global questions of special affine geometry for surfaces in \mathbb{R}^3 were first studied by Blaschke (see Blaschke {1; v.2}). Santalo [1] used moving frames, and derived integral formulas, for the same purpose, and Hsiung and Shahin [1] do the same thing for hypersurfaces in higher dimension. Section 10 of Flanders' paper gives a formula for the special affine second fundamental form \mathbf{II} —more precisely, it gives a formula for $\mathbf{II}(c', c', c')$ for a curve c

in M (compare Blaschke {1; v.2, §46} and Hsiung and Shahin [1, §4]). Section 2 indicates one crucial difference between the cases $n = 2$ and $n \geq 3$: in the latter case, the special affine Codazzi-Mainardi equations are a consequence of the apolarity conditions; this essentially depends on the Bianchi identities (compare Blaschke {1; v.2; §§65,66}). In ordinary geometry there is a similar situation for $n \geq 4$ (see Eisenhart {1; Appendix 22}).

Gardner, R. B.

- [1] *Subscalar pairs of metrics and hypersurfaces with a non-degenerate second fundamental form*, J. Differential Geometry **6** (1972), 437–458 (MR 48 #9608).

Green, L. W.

- [1] *Auf Wiedersehensflächen*, Ann. of Math. (2) **78** (1963), 289–299 (MR 27 #5206).

Green, R. E. and Wu, H.-H.

- [1] *On the rigidity of punctured ovaloids*, Ann. of Math. (2) **94** (1971), 1–20 (MR 44 #7490); *II*, J. Differential Geometry **6** (1972), 459–472 (MR 51 #11374).

Gromoll, D. and Meyer, W.

- [1] *On complete open manifolds of positive curvature*, Ann. of Math. (2) **90** (1969), 75–90 (MR 40 #854).

Gromov, M. L. and Rokhlin, V. A.

- [1] *Embeddings and immersions in Riemannian geometry*, Russian Math. Surveys **25** (1970), no. 5, 1–57 (MR 44 #7571).

Grove, V. G.

- [1] *On closed convex surfaces*, Proc. Amer. Math. Soc. **8** (1957), 777–786 (MR 19 167).

Hamburger, H.

- [1] *Beweis einer Carathéodoryschen Vermutung. Teil I*, Ann. of Math. **41** (1940), 63–86 (MR 1 172); *II*, Acta Math. **73** (1941), 175–228; *III*, Acta Math. **73** (1941), 229–332 (MR 3 310).

Hartman, P. and Nirenberg, L.

- [1] *On spherical image maps whose Jacobians do not change sign*, Amer. J. Math. **81** (1959), 901–920 (MR 23 #A4106).

Hayden, H. A.

- [1] *Asymptotic lines in a V_m in a V_n* , Proc. London Math. Soc (2) **33** (1931–32), 22–31.

Heinz, E.

- [1] *On Weyl's embedding problem*, J. Math. Mech. **11** (1962), 421–454 (MR 25 #2565).

Hellwig, G.

- [1] *Über die Verbiegbarkeit von Flächenstücken mit positiver Gauzscher Krümmung*, Arch. Math. (Basel) **6** (1955), 243–249 (MR 16 1148).

Hermann, R.

- [1] *The second variation for minimal submanifolds*, J. Math. Mech. **16** (1966), 473–491 (MR 34 #8348).

Chapter IV.9, Addendum 4 was based on the presentation in this paper.

Hilbert, D.

- [1] *Ueber Flächen von constanter Gauzscher Krümmung*, Trans. Amer. Math. Soc. **2** (1901), 87–99.

Hoesli, R. J.

- [1] *Spezielle Flächen mit Flachpunkten und ihre lokale Verbiegbarkeit*, Composito Math. **8** (1950), 113–141 (MR 12 357).

Holmgren, E.

- [1] *Sur les surfaces à courbure constante négative*, C. R. Acad. Sci. Paris Ser. A-B **134** (1902), 740–743.

Hopf, H.

- [1] *Über Flächen mit einer Relation Zwischen den Hauptkrümmungen*, Math. Nachr. **4** (1951), 232–249 (MR 12 634).
- [2] *Über die Curvatura integra geschlossener Hyperflächen*, Math. Ann. **95** (1925), 340–367.

Hopf, H. and Schilt, H.

- [1] *Über Isometrie und stetig Verbiegung von Flächen*, Math. Ann. **116** (1939), 58–75.

Hopf, H. and Voss, K.

- [1] *Ein Satz aus der Flächentheorie im Grossen*, Arch. Math. (Basel) **3** (1952), 187–192 (MR 14 #583).

This is the source for Theorem V.12-17; it relates this result to another, apparently quite different one.

Horn, R. A.

- [1] *On Fenchel's theorem*, Amer. Math. Monthly **78** (1971), 380–381 (MR 44 #2142).

Hsiung, C.-C and Shahin, J. K.

- [1] *Affine differential geometry of closed hypersurfaces*, Proc. London Math. Soc. (3) **17** (1967), 715–735 (MR 36 #2069).

Jacobowitz, H.

- [1] *Local isometric embeddings of surfaces into Euclidean four space*, Indiana Univ. Math. J. **21** (1971/72), 249–254 (MR 46 #6247).
- [2] *Extending isometric embeddings*, J. Differential Geometry **9** (1974), 291–307 (MR 51 #13942).

This paper, together with helpful hints from the author, was the source for the proof of Theorem V.11-9.

Jacobowitz, H. and Moore, J. D.

- [1] *The Cartan-Janet theorem for conformal embeddings*, Indiana Univ. Math. J. **23** (1973), 187–203 (MR 47 #5757).

This paper considers the problem of conformal embedding from two different approaches—the first using the methods to be found in the proof of Theorem V.11-9, and the second using the methods of Chapter V.11, Addendum. There is a long discussion of differential systems which should be useful for reading Cartan {4}.

Jessen, B.

see Fenchel, W.

Kapouleas, N.

- [1] *Compact constant mean curvature surfaces in Euclidean three-space*, J. Differential Geometry **33** (1991), no. 3., 683–715 (MR 93a:53007b).
- [2] *Constant mean curvature surfaces constructed by fusing Wente tori*, Proc. Nat. Acad. Sci. U.S.A. **89** (1992), no. 12, 5695–5698 (MR 93h:53011).

Klein, F.

- [1] *Vergleichende Betrachtungen über neuere geometrische Forschungen*, Math. Ann. **43** (1893), 63–100.

Also in volume 1, pp. 460–497 of *Klein's Gesammelte Mathematische Abhandlungen*, 3 vols., Springer, Berlin, 1921–23, and in English translation in Bull. N. Y. Math. Soc. **2** (1892), pg. 215.

Klotz, T.

- [1] *On G. Bol's proof of Carathéodory's conjecture*, Comm. Pure Appl. Math. **12** (1959), 277–311 (MR 22 #11352).

Klotz Milnor, T.

- [2] *Efimov's theorem about complete immersed surfaces of negative curvature*, Advances in Math. **8** (1972), 474–543 (MR 46 #835).

Knebelman, M. S.

- [1] *Contact transformations*, Ann. of Math. (2) **39** (1938), 507–515.

Kuiper, N. H.

- [1] *On C^1 -isometric imbeddings. I, II*, Nederl. Akad. Wetensch. Proc. Ser. A. **58** = Indag. Math. **17** (1955), 545–556, 683–689 (MR 17 782).

The results of this paper are strengthened somewhat in the next.

- [2] *Isometric and short imbeddings*, Nederl. Akad. Wetensch. Proc. Ser. A. **62** = Indag. Math. **21** (1959), 11–25 (MR 20 #7316).
- [3] *On surfaces in euclidean three-space*, Bull. Soc. Math. Belg. **12** (1960), 5–22 (MR 23 #A609).

This paper is the source for the arguments on pp. III.280–286.

Lashof, R. K.

see Chern, S.-S.

Lawson, H. B.

- [1] *Complete minimal surfaces in S^3* , Ann. of Math. (2) **92** (1970), 335–374 (MR 42 #5170).

Lax, P.

see Courant, R.

Levi, E. E.

- [1] *Sulla deformazione delle superficie flessibili ed inestendibili*, Atti Acad. Torino **43** (1907–1908), 292–302.

Lewy, H.

- [1] *An example of a smooth linear partial differential equation without solution*, Ann. of Math. (2) **66** (1957), 155–158 (MR 19 551).
- [2] *On the existence of a closed surface realizing a given Riemannian metric*, Proc. Nat. Acad. Sci. U.S.A. **24** (1938), 104–106.
- [3] *On differential geometry in the large I (Minkowski's problem)*, Trans. Amer. Math. Soc. **43** (1938), 258–270.
- [4] *Über das Anfangswertproblem einer hyperbolischen nichtlinearen partiellen Differentialgleichung zweiter Ordnung mit zwei unabhängigen Veränderlichen*, Math. Ann. **98** (1927), 179–191.
- [5] *Neuer Beweis des analytischen Charakters der Lösungen elliptischer Differentialgleichungen*, Math. Ann. **101** (1929), 609–619.

The last two papers are the source for Chapter V.10, sections 8 and 9.

Liebmann, H.

- [1] *Ein Satz über endliche einfach zusammenhängende Flächenstücke negative Krümmung*, Berichte über die Verhandlungen der Königlich Sächsischen Gessellschaft

der Wissenschaften zu Leipzig, Mathematisch-physische classe 52 (1900), 28–36.

- [2] *Die Verbiegung von geschlossenen und offenen Flächen positiver Krümmung*, Bayer. Akad. Wiss. Math.-Physik Kl. S.-B. (Munich, 1919), 267–291.
- [3] *Ueber die Verbiegung der geschlossenen Flächen positiver Krümmung*, Math. Ann. 53 (1900), 81–112.

This paper gives a geometric proof of the infinitesimal rigidity of convex surfaces, as well as an early proof that the sphere is unwarpable (Theorem III.5-2).

Lima, E.

see Carmo, M. do

Massey, W. S.

- [1] *Surfaces of Gaussian curvature zero in Euclidean 3 space*, Tôhoku Math. J. (2) 14 (1962), 73–79 (MR 25 #2527).

This paper gives the first of our three proofs of Theorem III.5-9. For other formulations of the theorem, see Hartman and Nirenberg [1] and Pogorelov [3], [4].

McCleary, J.

- [1] *On Jacobi's remarkable curve theorem*, Historia Math. 21 (1994), 377–385.

Meyer, W.

see Gromoll, D.

Moore, J. D.

- [1] *Isometric immersions of Riemannian products*, J. Differential Geometry 5 (1971), 159–168 (MR 46 #6249).
- [2] *Isometric immersions of space forms in space forms*, Pacific J. Math. 40 (1972), 157–166 (MR 46 #4442).

This is the source for Problem V.11-2.

see also Jacobowitz, H.

Nash, J.

- [1] *C^1 isometric imbeddings*, Ann. of Math. (2) 60 (1954), 383–396 (MR 16 515).
- [2] *The imbedding problem for Riemannian manifolds*, Ann. of Math. (2) 63 (1956), 20–63 (MR 17 782).

Nirenberg, L.

- [1] *The Weyl and Minkowski problems in differential geometry in the large*, Comm. Pure Appl. Math. 6 (1953), 337–394 (MR 15 347).

- [2] *Rigidity of a class of closed surfaces*, Nonlinear Problems (Proc. Sympos., Madison, Wis., 1962), pp. 177–193. Univ. of Wisconsin Press, Madison, Wis., 1963 (MR 27 #697).

see also Hartman, P.

Nitsche, J. C. C.

- [1] *Elementary proof of Bernstein's theorem on minimal surfaces*, Ann. of Math. (2) **66** (1957), 543–544 (MR 19 878).

O'Neil, B.

- [1] *Isometric immersions which preserve curvature operators*, Proc. Amer. Math. Soc. **13** (1962), 759–763 (MR 26 #721).

Olowjanischnikow, S. P.

- [1] *On the bending of infinite convex surfaces*, Mat. Sb. **18** (60) (1946), 429–440 [Russian. English summary] (MR 8 169).

Ostrowski, A.

- [1] *Un'applicazione dell'integrale di Stieltjes all'analisi elementare delle curve piane*, Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat. (8) **18** (1955), 373–375 (MR 17 780).
- [2] *Über die Verbindbarkeit von Linien- und Krümmungselementen durch monoton gekrümmte Kurvenbögen*, Enseignement Math. (2) **2** (1956), 277–292 (MR 19 58).

Palais, R.

- [1] *A topological Gauss-Bonnet theorem*, J. Differential Geometry **13** (1978), 385–398 (MR 82b:58002).

Pick, G.

- [1] *Über affine Geometrie IV: Differentialinvarianten der Flächen gegenüber affinen Transformationen*, Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-physische Klasse **69** (1917), 107–136.

Pogorelov, A. W.

- [1] *Isometric transformation of punctured convex surfaces*, Dokl. Akad. Nauk SSSR **137** (1961), 1307–1308 = Soviet Math. **2**, 475–476 (MR 25 #1520).
- [2] *On the rigidity of general infinite convex surfaces with integral curvature 2π* , Dokl. Akad. Nauk SSSR **106** (1956), 19–20 [Russian] (MR 17 888).
- [3] *Continuous maps of bounded variations*, Dokl. Akad. Nauk SSSR **111** (1956), 757–759 [Russian] (MR 19 309).

- [4] *Extensions of the theorem of Gauss on spherical representation to the case of surfaces of bounded extrinsic curvature*, Dokl. Akad. Nauk SSSR **111** (1956), 945–947 [Russian] (MR 19 309).

Rembs, E.

- [1] *Zur Verbiegung von Flächen im Grossen*, Math. Z. **56** (1952), 271–279 (MR 14 901).

Rokhlin, V. A.

see Gromov, M. L.

Ryan, P. J.

- [1] *Homogeneity and some curvature conditions for hypersurfaces*, Tôhoku Math. J. (2) **21** (1969), 363–388 (MR 40 #6458).

Saban, G.

- [1] *Nuove caratterizzazioni della sfera*, Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat. **25** (1958), 457–464 (MR 21 #5967).

Sacksteder, R.

- [1] *On hypersurfaces with no negative sectional curvatures*, Amer. J. Math. **82** (1960), 609–630 (MR 22 #7087).

Santalo, L. A.

- [1] *Geometria diferencial afín y cuerpos convexos*, Math. Notae **16** (1957), 20–42 (MR 20 #6713).

Scherrer, W.

- [1] *Eine Kennzeichnung der Kugel*, Vierteljschr. Naturforsch. Ges. Zürich **85** Beiblatt (Festschrift Rudolf Feuter) 1940, 40–46 (MR 3 89).

Schilt, H.

- [1] *Über die isolierten Nullstellen der Flächenkrümmung und einige Verbiegbarkeitssätze*, Composito Math. **5** (1937), 239–283.

see also Hopf, H.

Shahin, J. K.

see Hsiung, C.-C.

Stoker, J. J.

- [1] *Über die Gestalt der positiv gekrümmten offenen Flächen im dreidimensionalen Raume*, Composito Math. **3** (1936), 55–89.

Süss, W.

- [1] *Zur relativen Differentialgeometrie III: Über Relativ-Minimalflächen und Verbiegung*, Japan J. Math. **4** (1928), 203–207.

Vermeil, H.

- [1] *Bestimmung einer quadratischen Differentialform aus der Riemannschen und den Christoffelschen Differentialinvarianten mit Hilfe von Normalkoordinaten*, Math. Ann. **79** (1918), 289–312.

Volkov, Ju. A. and Vladimirova, S. M.

- [1] *Isometric immersions of the Euclidean plane in Lobachevskii space*, Mat. Zametki **10** (1971), 327–332 = Math. Notes **10** (1971), 619–622 (MR 45 #2624).

Voss, K.

- [1] *Eine Bemerkung über Totalkrümmung geschlossener Raumkurven*, Arch. Math. (Basel) **6** (1955) 259–263 (MR 17 75).
see also Hopf, H.

Walden, R.

- [1] *Eindeutigkeitssätze für II-isometrische Eiflächen*, Math. Z. **120** (1971), 143–147 (MR 44 #2182).

Warner, F. W.

see Carmo, M. do

Weil, A.

see Allendoerfer, C. B.

Weiner, J. L.

- [1] *Closed curves of constant torsion*, Arch. Math. (Bassel) **25** (1974), 313–317 (MR 49 #11437); *II*, Proc. Amer. Math. Soc **67** (1977), no. 2, 306–308 (MR 57 #1370).

Wente, H. C.

- [1] *Counterexample to a conjecture of H. Hopf*, Pacific J. Math. **121** (1986), no. 1, 193–243 (MR 87d:53013).

Wu, H.-H.

see Green, R. E.

Wunderlich, W.

- [1] *Über ein abwickelbares Möbiusband*, Montasch. Math. **66** (1962), 276–289 (MR 26 #680).

NOTATION INDEX

CHAPTER 10

$c_{k+1}(p, X_1, \dots, X_k)$	114
$c_{k+1}(W)$	114
$\mathcal{E}(p, X_1, \dots, X_k)$	113
$\mathcal{E}(W)$	113
F_x, F_y, F_u, F_p, F_q	3
\mathcal{M}_k	113
p, q	3, 38
r, s, t	38

CHAPTER 11

$\mu(p)$	141
$\nu(p)$	141
$\tilde{\Phi}$	137

CHAPTER 12

det	182
h	184
$t(p)$	170
$\omega \times \eta$	181
$\omega \bullet \eta$	181

CHAPTER 13

$\text{Ad}(a)$	308
$\text{Ad}(a)^*$	311
A_n	300
A^*	358
$A \oplus B$	290
a_P	285
$C(\xi)$	296, 297
$c_k(\xi)$	364
$c_{n,k}$	364
$c(\xi)$	364
DM	381
$D_{sr}f$	321
\det	325
\det_{r_1, \dots, r_n}	322, 373
\det_{s_1, \dots, s_n}^*	373
e_n	344

$e_n(\xi)$	345
e_{ri}	320
\exp	358
$e(\xi)$	345
$F(\xi)$	266
\tilde{f}	377
fg	354
$f_i(M)$	330
$\tilde{f}_k(M)$	362
$f(\zeta)$	341
$f(\psi)$	363
$f(\Omega)$	354
$f^*\xi$	268
$\text{GL}(n, k)$	372
$G_n(\mathbb{C}^N)$	360
$G_n(\mathbb{R}^N)$	273
$\tilde{G}_n(\mathbb{R}^N)$	281
$G_n(\mathbb{R}^\infty)$	280
$\tilde{G}_n(\mathbb{R}^\infty)$	304
h	359
\mathfrak{h}'	313
\mathfrak{h}^\perp	314
L_A	346
L_a	305
\mathbf{L}_A	346
\mathbf{L}_a	305
$\langle M, P \rangle$	308, 363
$\text{O}(E)$	294
$\text{O}(n) \times \text{O}(N - n)$	275
$\text{O}(\xi)$	272
\mathcal{P}	285
$\mathcal{P}f$	353
$\text{Pf}(A)$	285, 289
$\mathcal{P}^k(\mathfrak{g})$	353
\mathcal{P}'	287
$\mathcal{P}(\mathfrak{g})$	353
$p_k(\xi)$	345
$p_{n,k}$	344
$p_{n,k}(\xi)$	345

$p(\xi)$	351	$\gamma^n(\mathbb{C}^N)$	360
R_A	346	$\gamma^n(\mathbb{R}^N)$	276
\mathbf{R}_A	346	$(\tilde{\gamma}^n(\mathbb{R}^N), \mu)$	283
δf	354	Δ_{sr}	322
$\delta^k(\mathfrak{g})$	353	$\Delta_{s_2 r_2} \Delta_{s_1 r_1}$	322
$\mathrm{SL}(n, \mathbb{C})$	376	$\varepsilon^{j_1 \dots j_n}$	284
$\mathrm{SL}(n, k)$	373	ε_{rs}	373
$\mathfrak{sl}(n, \mathbb{C})$	376	$\varepsilon(P)$	285
$\mathrm{SO}(E)$	294	$\zeta_{i_1 i_2}$	339
$\mathrm{SU}(n)$	376	l_{rs}	319
$\mathfrak{su}(n)$	376	$\bar{\xi}$	366
T^*	361	$\xi_{\mathbb{R}}$	360
$\mathrm{U}(n)$	358	$\xi_1 \oplus \xi_2$	267
$U(\xi)$	291, 360	π_1, π_2	346
$\mathfrak{u}(n)$	358	σ_i	317
\bar{V}	366	σ^n	308
$V_{\mathbb{C}}$	365	$\tau: \tilde{G}_n(\mathbb{R}^N) \rightarrow$	
$V_n(\mathbb{R}^N)$	273	$G_n(\mathbb{R}^N)$	282
$V_n^{\mathrm{O}}(\mathbb{R}^N)$	273	ϕ_{α}^{β}	337
$V_{\mathbb{R}}$	358	$\tilde{\phi}_{\alpha}^{\beta}$	346
$ v $	357	$\chi(M)$	381
w_{ξ}	355	$\chi(\xi)$	291
X_{α}^{β}	337	ψ_{ij}	363
$\tilde{\alpha}_{jk}$	359	Ωf	325
$\alpha_{N, N'}: \tilde{G}_n(\mathbb{R}^N) \rightarrow$		Ω_j^i	349
$\tilde{G}_n(\mathbb{R}^{N'})$	302	$\Omega^k(\mathfrak{g}/\mathfrak{h})$	313
$\alpha: G_n(\mathbb{C}^N) \rightarrow$		$\Omega^k(\mathfrak{g}/\mathfrak{h})^H$	313
$G_n(\mathbb{C}^M)$	360	$\Omega^k(\mathfrak{h}^{\perp})^H$	314
$\alpha: G_n(\mathbb{R}^N) \rightarrow$		ω_j^i	346
$G_n(\mathbb{R}^M)$	277	ϖ	346
$\alpha: \tilde{G}_n(\mathbb{R}^N) \rightarrow$		\bullet	366
$\tilde{G}_n(\mathbb{R}^M)$	283	$<$	326
γ^n	280		

INDEX

- Adjoint, of linear transformation, 361
- $\text{Ad}(G)$ -invariant, 353, 354
- $\text{Ad}(H)$ -invariant, 311
- Alexandrov, A. D., 155, 156, 211
- Algebraic description of $H^k(G/H)$, 313
- Algebraic identities, principle of extension of, 286
- Algebraic inequalities, principle of irrelevance of, 375
- Allendoerfer, C. B., 251, 264, 265
- Ampère, A.-M., *see* Monge-Ampère equations
- Analyticity of elliptic solutions of second order equations in 2 variables, 98–109
- Annihilates a Lie subalgebra, 311
- Asymptotic curves in imbedding and rigidity problems, 218–223
- Base curve, 19
- Beez, R., 170
- Beltrami equation, 53
- Bendable, 170
 - infinitesimally, 173
- Bending, 170
 - infinitesimal, 173
 - of rotation surfaces, 253
 - through imbeddings, 170
 - through immersions, 170
- Blaschke, W., 213
- Blaschke, integral formula of, 188, 199
- Bonnesen, T., 200
- Bonnet, O., *see* Gauss-Bonnet theorem
- Boundary value problem, 67
- Brunn-Minkowski inequality, 200
- Bundle
 - complex, 356
 - conjugate, 366
 - induced, 268
 - map, 265–266, 356
 - oriented, 281
 - sphere, 268
 - universal, 280
- Burstin, C., 157
- Burstin-Janet-Cartan Theorem, 149
- Canonical form
 - for $\mathfrak{o}(n)$, 287, 331
 - for $\mathfrak{u}(n)$, 361
- Capacitance, heat, 63
- Capelli identities, 325
- Cartan, É., 138, 140, 149, 157, 170, 208; *see also* Burstin-Janet-Cartan Theorem, Cartan-Kähler theorem
- Cartan-Kähler theorem, 111, 121
- Cauchy problem
 - for a general first order PDE, 25, 28
 - for a higher order PDE, 29–35
 - reduction to Cauchy problem for a system of first order PDE's, 36
 - for a linear first order PDE, 8
 - for a quasi-linear first order PDE, 13
- Cauchy-Kowalewski theorem, 38–46
- Cayley Ω -process, 325
- Central projection, 235
- Characteristic
 - for given initial conditions for a quasi-linear first order PDE, 13
 - for given initial data for a general first order PDE, 26
 - for given initial data for a hyperbolic second order PDE in 2 variables, 97
 - for given initial data for a second order PDE, 35
- Characteristic classes, 302ff.; *see also* Chern class, Euler class, Pontryagin class
- Characteristic curve
 - of a function, for a hyperbolic system in 2 variables in diagonal form, 79
 - of a linear first order PDE, 4
 - of a quasi-linear first order PDE, 10, 26

- Characteristic curve (*continued*)
 - of a solution of a first order PDE, 15, 18
- Characteristic strip, 20
- Characteristic vector field
 - of a linear first order PDE, 4
 - of a quasi-linear first order PDE, 10, 26
 - of a semi-linear hyperbolic system in 2 variables, 74
 - of a solution of general first order PDE, 18
- Chern, S.-S., 200, 204, 208, 249, 265, 301, 380, 381
- Chern class, 364, 370
 - total, 364
- Christoffel, E. B., 206
- Christoffel's Theorem, 204, 207
- Classification of semi-linear second order PDE's, 47–59
- Classifying space, 280
- Cohn-Vossen, S., 200, 210, 213, 213, 214; *see also* Hilbert and Cohn-Vossen
- Cohn-Vossen's Theorem, 192, 212
- Complete convex surfaces, rigidity of, 211–212
- Complex
 - Grassmannian, 360
 - orthogonal group, 357
 - vector bundle, 356
- Conductivity, heat, 64
- Cone, Monge, 15
- Conjugate
 - bundle, 366
 - linear, 357
- Constant curvature manifolds
 - in manifolds of greater constant curvature, 140
 - Pontryagin classes of, 351
- Convex complete surfaces, rigidity of, 211–212
- Courant, R., 79
- Covering homotopy theorem, 271
- Curvature, *see* Constant curvature manifolds
- Curve
 - base, 19
 - characteristic, *see* Characteristic curve
 - initial, *see* Initial curve
- Darboux, G., 144
- Darboux equation, 145, 217ff.
- Degree, total, 326
- Density, 61
- Dependence, domain of, 70
- Depends on q variables, 139
- Derivative, normal, 32
- Differential equations, partial, *see* Partial differential equations
- Differential ideal (or system), 112
 - integral element of, 112
 - integral submanifold of, 112
- Dirichlet problem, 68
- Dolbeault-Lemoine, S., 246
- Domain of dependence, 70
- Double of a manifold, 381
- Efimov, N. V., 235
- Elementary symmetric functions, 317
- Elliptic
 - semi-linear PDE, 49
 - solution of a second order PDE, 57
 - solution of a second order PDE
 - analyticity of, in 2 variables, 98–109
- Energy, heat, 63
- Equivalence, 265, 266
- Euler class, 291, 350, 370
- Euler's theorem, 320
- Evaluations, 373
- Extension of algebraic identities, principle of, 286
- Exteriorly orthogonal, 137

- Fenchel, W., 156, 200, 264
 Firey, W.J., 208
 First normal space, 247
 First order PDE's, 3
 general, 13
 Cauchy problem for, 25, 28
 characteristic curve of a solution, 15, 18
 characteristic strip of, 20, 27
 characteristic vector field of a solution, 18
 initial curve for
 characteristic, 26
 free, 25
 initial data for, 22, 28
 initial manifold for, free, 28
 linear, 4
 Cauchy problem for, 8
 characteristic curve of, 4
 characteristic vector field of, 4
 initial condition for, 8
 initial curve for, free, 8
 quasi-linear, 9
 Cauchy problem for, 13
 characteristic curve of, 10, 26
 characteristic vector field of, 10, 26
 initial conditions for, 12, 27
 initial curve for
 characteristic, 13
 free, 12
 initial manifold for, free, 27
 systems of, 36
 Cauchy-Kowalewski theorem, 38–46
 Form, second fundamental, 208
 Free
 initial curve
 for a general first order PDE, 25
 for a linear first order PDE, 8
 for a quasi-linear first order PDE, 12
 initial manifold
 for a higher order PDE, 29–35
 for a quasi-linear first order PDE, 27
 Fundamental form, second, 208
 Gardner, R. B., 209
 Gauss-Bonnet (-Chern) theorem, 263–265, 301, 380ff.
 Geodesic parallels, 224
 Good basis for a regular integral element, 114
 Grassmannian manifold, 273
 complex, 360
 oriented, 281
 Green, R. E., 214
 Gromov, M. L., 156
 Grove, V. G., 208
 Harmonic function, 71
 Heat
 capacitance, 63
 conductivity, 64
 energy, 63
 equation, 65ff.
 specific, 63
 Heinz, E., 155
 Hellwig, G., 213, 234
 Herglotz integral formula, 194, 197, 199
 Hermitian
 inner product, 357
 metric, 360
 Higher order PDE's, 29–35
 Hilbert, D., 213, 214
 Hoesli, R.J., 235
 Homogeneous
 ideal, 112
 polynomial function, 319
 space, 304
 Homotopy covering theorem, 271
 Hopf, E., 127
 Hopf, H., 206, 234, 263
 Hopf's Theorem, 127, 200
 Hyperbolic
 second order equation in 2 variables, 81–97
 semi-linear PDE, 49

- Hyperbolic (*continued*)
 - solution of a second order PDE, 57
 - system in 2 variables, 72–80
 - C^k , 73
 - semi-linear, 73
- Ideal
 - differential, 112
 - homogeneous, 112
- Identities, principle of extension of
 - algebraic, 286
- Imbedding(s)
 - bending of, 170
 - bending through, 170
 - isometric, 133ff.
- Immersion, bending through, 170
- Index
 - of nullity, 141
 - of relative nullity, 141
- Induced bundle, 268
- Inequalities, principle of irrelevance of
 - algebraic, 375
- Infinitesimal
 - bending, 173
 - of rotation surfaces, 253
 - rotation field, 175
- Infinitesimally
 - bendable, 173
 - rigid, 173
- Initial-boundary value problem, 67
- Initial condition, 2
 - for a higher order PDE, 29–35
 - for a linear first order PDE, 8
 - for a quasi-linear first order PDE, 12, 27
- Initial curve
 - for a general first order PDE
 - characteristic, 26
 - free, 25
 - for a linear first order PDE, free, 8
 - for a quasi-linear first order PDE
 - characteristic, 13
 - free, 12
- Initial data
 - for a first order PDE, 22
 - for a higher order PDE, 29–35
- Initial manifold
 - free for a general first order PDE, 28
 - free for a higher order PDE, 29–35
 - free for a quasi-linear first order PDE, 27
- Inner product, Hermitian, 357
- Integrability conditions, 1, 112
- Integral element of a differential
 - system, 112
 - regular, 113
 - good basis for, 114
- Integral formula
 - of Blaschke, 188, 199
 - of Herglotz, 194, 197, 199
- Integral submanifold of a differential
 - ideal, 112
- Invariant
 - $\text{Ad}(G)$, 354
 - $\text{Ad}(H)$, 311
 - differential form, 308
 - under a subgroup of $\text{GL}(n, k)$, 372
 - under adjoint action of $\text{O}(n)$ and $\text{SO}(n)$, 331
 - under adjoint action of $\text{U}(n)$, 362
 - under $\text{O}(n)$, 318
- Invariant theory, 317ff.
 - first main theorem for $\text{O}(n)$, 327ff.
 - first main theorem for $\text{U}(n)$, 361, 379
 - for unitary group, 372ff.
- Irrelevance of algebraic inequalities,
 - principle of, 375
- Isometric imbeddings, 133ff.
- Isothermal
 - coordinates, 52
 - surface, 261
- Janet, M., 149, 157
- Jessen, B., 156

- Kähler, E., *see* Cartan-Kähler theorem
 Killing, W., 170
 Kowalewski, S., *see* Cauchy-Kowalewski theorem
 Kuiper, H. H., 156
- Laplace equation, 66
 Lax, P., 79
 Length, preserve up to first order, 173
 Levi, E. E., 229, 230, 231
 Lewy, H., 46, 155, 156
 Liebmann, H., 199, 213
 Linear first order PDE, *see* First order PDE
- Majorants, method of, 43
 Manifold
 initial, *see* Initial manifold
 strip, *see* Strip
 Manifolds of constant curvature,
 see Constant curvature manifolds
- Map
 bundle, 265, 266, 356
 natural between Grassmannians,
 277, 360
- Maximum principle, 126ff.
 Metric, Hermitian, 360
 Milnor, J. W., 356, 370
 Minkowski's
 formulas, 185
 problem, 156, 200
 theorem, 200
 generalized, 207
 see also Brunn-Minkowski inequality
- Monge cone, 15
 Monge-Ampère equations, 97, 145,
 220
 Moore, J. C., 252
- Nash, J., 156
 Natural
 classes, 302
 map between Grassmannians, 277,
 360
- Nirenberg, L., 155, 156, 211
 Non-characteristic, 8
 Non-compact surfaces, complete
 convex, 211
 Non-convex surfaces, 209
 Non-degenerate, 148
 Non-trivial
 bending, 170
 infinitesimal bending, 173
- Normal
 derivative, 32
 form for a semi-linear second order
 PDE, 50
 linear transformation, 361
 space, first, 247
- Nullity
 index of, 141
 relative index of, 141
- Olowjanischnikow, S. P., 211
 Orientation
 of direct sum of vector spaces, 281
 of vector bundle, 281
 of vector space, 280
- Oriented
 Grassmannian manifold, 281
 vector bundle, 281
 vector space, 281
- Orthogonal
 exteriorly, 137
 group, complex, 357
- Orthonormal, with respect to Hermitian inner product, 357

- Parabolic
 - semi-linear PDE, 49
 - solution of a second order PDE, 57
- Parallels, geodesic, 224
- Part, principal, 47
- Partial differential equations, 1
 - first order, *see* First order PDE's
 - higher order, *see* Higher order PDE's and second order PDE's
 - linear first order, *see* First order PDE's
 - quasi-linear first order, *see* First order PDE's
 - second order, *see* Second order PDE's
 - system of, 1
 - integrability conditions for, 1, 112
 - overdetermined, 1
- PDE, *see* Partial differential equations
- Pfaffian, 285
- Physics, prototypical second order equations of, 59–72
- Pogorelov, A. W., 155, 212, 213, 214, 235
- Polar space, 113
- Polarization, 321
- Polynomial function, 319, 353, 360, 372
- Pontryagin class, 345, 370
 - of constant curvature manifolds, 351
 - total, 351
- Preserve lengths up to first order, 173
- Principal
 - bundle
 - map, 266
 - of frames, 266
 - trivial, 266
 - curvatures, radii of, 204
 - part, 47
- Projection, central, 235
- Quasi-linear first order PDE, *see* First order PDE's
- Radii of principal curvature, 204
- Rank, 326
- Regular integral element of a differential system, 113
 - good basis for, 114
- Relative nullity, index of, 141
- Rembs, E., 210
- Rigid, 173
 - infinitesimally, 173
- Rigidity of complete convex surfaces, 211–212
- Rokhlin, V. A., 156
- Rotation field, infinitesimal, 175
- Rotation surfaces, infinitesimal bendings of, 253
- Russian school of differential geometry, 155, 156, 212, 213
- Schilt, H., 231, 233, 234
- Schwartzschild metric, 133
- Second fundamental form, 208
- Second order PDE's
 - analyticity of elliptic solutions in 2 variables, 98–109
 - characteristic initial manifold for, 35
 - classification of semi-linear, 47–59
 - elliptic solution of, 57
 - analyticity of in 2 variables, 98–109
 - free initial manifold for, 29–35
 - hyperbolic solution of, 57
 - in 2 variables, 81–97
 - Monge-Ampère, 97
 - prototypical equations of physics, 59–72
 - semi-linear, 47
 - classification of, 47–59
 - elliptic, 49
 - hyperbolic, 49
 - normal forms of, 50
 - parabolic, 49
 - principal part of, 47
- Section, 305
- Semi-linear, *see* Second order PDE's

- Skew-Hermitian matrices, canonical form for, 361
- Skew-symmetric matrices, canonical form for, 287, 331
- Small vibrations, 62
- Specific heat, 63
- Spectral theorem, 361
- Sphere bundle, 268
- Star-shaped
 - subset of S^{n+1} , 240
 - surfaces of constant mean curvature, 186
- Stasheff, J. D., 356, 370
- Stiefel manifold, 273
- Stoker, J. J., 199
- String, vibrating, 59ff.
- Strip, 20
 - characteristic, 20, 27
 - condition, 20
 - manifold condition, 28
- Submanifold, integral of a differential ideal, 112
- Subsonic flow, 57
- Supersonic flow, 57
- Support function, 184
- Symmetric function, 317
 - elementary, 317
- System
 - differential, 112
 - hyperbolic in 2 variables, *see* Hyperbolic
 - integrability conditions for, 1, 112
 - of first order PDE's, 36
 - overdetermined, 1
- Temperature, 62
- Tension, 60, 61
- Thom class, 291, 370
- Tompkins, C., 137
- Torus of revolution, 210
- Total Chern class, 364
- Total degree, 326
- Total Pontryagin class, 351
- Trivial
 - bending, 170
 - infinitesimal bending, 173
 - principal bundle, 266
- Type number
 - of hypersurface, 170
 - of linear transformations, 247
 - of matrices, 247
 - of submanifold, 247
- Unbendable, 170
- Uniquely determined, 171
- Universal bundle, 280
- Unwarpable, 171
- Variation vector field, 171
- Vector bundle, complex, 356
- Vector field, characteristic, *see* Characteristic vector field
- Vector field, variation, 171
- Vibrating string, 59ff.
- Voss, K., 206
- Voss, A., 231
- Walden, R., 208
- Warpable, 171
- Watson, G. H., 257
- Wave, 69
- Wave equation, 62, 68–71
- Weak topology, 280
- Weil, A., 265
- Weil homomorphism, 355
- Weyl, H., 155
- Weyl's problem, 155, 191
- Whitney product formula, 352, 365
- Whitney sum, 267
- Whittaker, E. T., 257
- Wu, H., 214

Michael Spivak

Brandeis University

Calculus on Manifolds

A MODERN APPROACH TO CLASSICAL THEOREMS
OF ADVANCED CALCULUS



ADDISON-WESLEY PUBLISHING COMPANY

The Advanced Book Program

Reading, Massachusetts • Menlo Park, California • New York
Don Mills, Ontario • Wokingham, England • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid • San Juan • Paris
Seoul • Milan • Mexico City • Taipei

Calculus on Manifolds

A Modern Approach to Classical Theorems of Advanced Calculus

Copyright © 1965 by Addison-Wesley Publishing Company

All rights reserved

Library of Congress Card Catalog Number 66-10910

Manufactured in the United States of America

The manuscript was put into production on April 21, 1965;

this volume was published on October 26, 1965

ISBN 0-8053-9021-9

24 25 26 27 28-CRW-9998979695

Twenty-fourth printing, January 1995

Editors' Foreword

Mathematics has been expanding in all directions at a fabulous rate during the past half century. New fields have emerged, the diffusion into other disciplines has proceeded apace, and our knowledge of the classical areas has grown ever more profound. At the same time, one of the most striking trends in modern mathematics is the constantly increasing interrelationship between its various branches. Thus the present-day students of mathematics are faced with an immense mountain of material. In addition to the traditional areas of mathematics as presented in the traditional manner—and these presentations do abound—there are the new and often enlightening ways of looking at these traditional areas, and also the vast new areas teeming with potentialities. Much of this new material is scattered indigestibly throughout the research journals, and frequently coherently organized only in the minds or unpublished notes of the working mathematicians. And students desperately need to learn more and more of this material.

This series of brief topical booklets has been conceived as a possible means to tackle and hopefully to alleviate some of

these pedagogical problems. They are being written by active research mathematicians, who can look at the latest developments, who can use these developments to clarify and condense the required material, who know what ideas to underscore and what techniques to stress. We hope that they will also serve to present to the able undergraduate an introduction to contemporary research and problems in mathematics, and that they will be sufficiently informal that the personal tastes and attitudes of the leaders in modern mathematics will shine through clearly to the readers.

The area of differential geometry is one in which recent developments have effected great changes. That part of differential geometry centered about Stokes' Theorem, sometimes called the fundamental theorem of multivariate calculus, is traditionally taught in advanced calculus courses (second or third year) and is essential in engineering and physics as well as in several current and important branches of mathematics. However, the teaching of this material has been relatively little affected by these modern developments; so the mathematicians must relearn the material in graduate school, and other scientists are frequently altogether deprived of it. Dr. Spivak's book should be a help to those who wish to see Stokes' Theorem as the modern working mathematician sees it. A student with a good course in calculus and linear algebra behind him should find this book quite accessible.

Robert Gunning
Hugo Rossi

Princeton, New Jersey
Waltham, Massachusetts
August 1965

Preface

This little book is especially concerned with those portions of “advanced calculus” in which the subtlety of the concepts and methods makes rigor difficult to attain at an elementary level. The approach taken here uses elementary versions of modern methods found in sophisticated mathematics. The formal prerequisites include only a term of linear algebra, a nodding acquaintance with the notation of set theory, and a respectable first-year calculus course (one which at least mentions the least upper bound (sup) and greatest lower bound (inf) of a set of real numbers). Beyond this a certain (perhaps latent) rapport with abstract mathematics will be found almost essential.

The first half of the book covers that simple part of advanced calculus which generalizes elementary calculus to higher dimensions. Chapter 1 contains preliminaries, and Chapters 2 and 3 treat differentiation and integration.

The remainder of the book is devoted to the study of curves, surfaces, and higher-dimensional analogues. Here the modern and classical treatments pursue quite different routes; there are, of course, many points of contact, and a significant encounter

occurs in the last section. The very classical equation reproduced on the cover appears also as the last theorem of the book. This theorem (Stokes' Theorem) has had a curious history and has undergone a striking metamorphosis.

The first statement of the Theorem appears as a postscript to a letter, dated July 2, 1850, from Sir William Thomson (Lord Kelvin) to Stokes. It appeared publicly as question 8 on the Smith's Prize Examination for 1854. This competitive examination, which was taken annually by the best mathematics students at Cambridge University, was set from 1849 to 1882 by Professor Stokes; by the time of his death the result was known universally as Stokes' Theorem. At least three proofs were given by his contemporaries: Thomson published one, another appeared in Thomson and Tait's *Treatise on Natural Philosophy*, and Maxwell provided another in *Electricity and Magnetism* [13]. Since this time the name of Stokes has been applied to much more general results, which have figured so prominently in the development of certain parts of mathematics that Stokes' Theorem may be considered a case study in the value of generalization.

In this book there are three forms of Stokes' Theorem. The version known to Stokes appears in the last section, along with its inseparable companions, Green's Theorem and the Divergence Theorem. These three theorems, the classical theorems of the subtitle, are derived quite easily from a modern Stokes' Theorem which appears earlier in Chapter 5. What the classical theorems state for curves and surfaces, this theorem states for the higher-dimensional analogues (manifolds) which are studied thoroughly in the first part of Chapter 5. This study of manifolds, which could be justified solely on the basis of their importance in modern mathematics, actually involves no more effort than a careful study of curves and surfaces alone would require.

The reader probably suspects that the modern Stokes' Theorem is at least as difficult as the classical theorems derived from it. On the contrary, it is a very simple consequence of yet another version of Stokes' Theorem; this very abstract version is the final and main result of Chapter 4.

It is entirely reasonable to suppose that the difficulties so far avoided must be hidden here. Yet the proof of this theorem is, in the mathematician's sense, an utter triviality—a straightforward computation. On the other hand, even the statement of this triviality cannot be understood without a horde of difficult definitions from Chapter 4. There are good reasons why the theorems should all be easy and the definitions hard. As the evolution of Stokes' Theorem revealed, a single simple principle can masquerade as several difficult results; the proofs of many theorems involve merely stripping away the disguise. The definitions, on the other hand, serve a twofold purpose: they are rigorous replacements for vague notions, and machinery for elegant proofs. The first two sections of Chapter 4 define precisely, and prove the rules for manipulating, what are classically described as "expressions of the form" $P dx + Q dy + R dz$, or $P dx dy + Q dy dz + R dz dx$. Chains, defined in the third section, and partitions of unity (already introduced in Chapter 3) free our proofs from the necessity of chopping manifolds up into small pieces; they reduce questions about manifolds, where everything seems hard, to questions about Euclidean space, where everything is easy.

Concentrating the depth of a subject in the definitions is undeniably economical, but it is bound to produce some difficulties for the student. I hope the reader will be encouraged to learn Chapter 4 thoroughly by the assurance that the results will justify the effort: the classical theorems of the last section represent only a few, and by no means the most important, applications of Chapter 4; many others appear as problems, and further developments will be found by exploring the bibliography.

The problems and the bibliography both deserve a few words. Problems appear after every section and are numbered (like the theorems) within chapters. I have starred those problems whose results are used in the text, but this precaution should be unnecessary—the problems are the most important part of the book, and the reader should at least attempt them all. It was necessary to make the bibliography either very incomplete or unwieldy, since half the major

branches of mathematics could legitimately be recommended as reasonable continuations of the material in the book. I have tried to make it incomplete but tempting.

Many criticisms and suggestions were offered during the writing of this book. I am particularly grateful to Richard Palais, Hugo Rossi, Robert Seeley, and Charles Stenard for their many helpful comments.

I have used this printing as an opportunity to correct many misprints and minor errors pointed out to me by indulgent readers. In addition, the material following Theorem 3-11 has been completely revised and corrected. Other important changes, which could not be incorporated in the text without excessive alteration, are listed in the Addenda at the end of the book.

Michael Spivak

Waltham, Massachusetts
March 1968

Contents

Editors' Foreword, v

Preface, vii

1. Functions on Euclidean Space 1

NORM AND INNER PRODUCT, 1

SUBSETS OF EUCLIDEAN SPACE, 5

FUNCTIONS AND CONTINUITY, 11

2. Differentiation 15

BASIC DEFINITIONS, 15

BASIC THEOREMS, 19

PARTIAL DERIVATIVES, 25

DERIVATIVES, 30

INVERSE FUNCTIONS, 34

IMPLICIT FUNCTIONS, 40

NOTATION, 44

3. <i>Integration</i>	46
BASIC DEFINITIONS, 46	
MEASURE ZERO AND CONTENT ZERO, 50	
INTEGRABLE FUNCTIONS, 52	
FUBINI'S THEOREM, 56	
PARTITIONS OF UNITY, 63	
CHANGE OF VARIABLE, 67	
 4. <i>Integration on Chains</i>	 75
ALGEBRAIC PRELIMINARIES, 75	
FIELDS AND FORMS, 86	
GEOMETRIC PRELIMINARIES, 97	
THE FUNDAMENTAL THEOREM OF CALCULUS, 100	
 5. <i>Integration on Manifolds</i>	 109
MANIFOLDS, 109	
FIELDS AND FORMS ON MANIFOLDS, 115	
STOKES' THEOREM ON MANIFOLDS, 122	
THE VOLUME ELEMENT, 126	
THE CLASSICAL THEOREMS, 134	
 <i>Bibliography, 139</i>	
<i>Index, 141</i>	

Calculus on Manifolds

I

Functions on Euclidean Space

NORM AND INNER PRODUCT

Euclidean n -space \mathbf{R}^n is defined as the set of all n -tuples (x^1, \dots, x^n) of real numbers x^i (a “1-tuple of numbers” is just a number and $\mathbf{R}^1 = \mathbf{R}$, the set of all real numbers). An element of \mathbf{R}^n is often called a point in \mathbf{R}^n , and \mathbf{R}^1 , \mathbf{R}^2 , \mathbf{R}^3 are often called the line, the plane, and space, respectively. If x denotes an element of \mathbf{R}^n , then x is an n -tuple of numbers, the i th one of which is denoted x^i ; thus we can write

$$x = (x^1, \dots, x^n).$$

A point in \mathbf{R}^n is frequently also called a vector in \mathbf{R}^n , because \mathbf{R}^n , with $x + y = (x^1 + y^1, \dots, x^n + y^n)$ and $ax = (ax^1, \dots, ax^n)$, as operations, is a vector space (over the real numbers, of dimension n). In this vector space there is the notion of the length of a vector x , usually called the **norm** $|x|$ of x and defined by $|x| = \sqrt{(x^1)^2 + \dots + (x^n)^2}$. If $n = 1$, then $|x|$ is the usual absolute value of x . The relation between the norm and the vector space structure of \mathbf{R}^n is very important.

1-1 Theorem. If $x, y \in \mathbf{R}^n$ and $a \in \mathbf{R}$, then

- (1) $|x| \geq 0$, and $|x| = 0$ if and only if $x = 0$.
- (2) $|\sum_{i=1}^n x^i y^i| \leq |x| \cdot |y|$; equality holds if and only if x and y are linearly dependent.
- (3) $|x + y| \leq |x| + |y|$.
- (4) $|ax| = |a| \cdot |x|$.

Proof

- (1) is left to the reader.
- (2) If x and y are linearly dependent, equality clearly holds.
If not, then $\lambda y - x \neq 0$ for all $\lambda \in \mathbf{R}$, so

$$\begin{aligned} 0 < |\lambda y - x|^2 &= \sum_{i=1}^n (\lambda y^i - x^i)^2 \\ &= \lambda^2 \sum_{i=1}^n (y^i)^2 - 2\lambda \sum_{i=1}^n x^i y^i + \sum_{i=1}^n (x^i)^2. \end{aligned}$$

Therefore the right side is a quadratic equation in λ with no real solution, and its discriminant must be negative. Thus

$$4 \left(\sum_{i=1}^n x^i y^i \right)^2 - 4 \sum_{i=1}^n (x^i)^2 \cdot \sum_{i=1}^n (y^i)^2 < 0.$$

- (3) $|x + y|^2 = \sum_{i=1}^n (x^i + y^i)^2$
 $= \sum_{i=1}^n (x^i)^2 + \sum_{i=1}^n (y^i)^2 + 2 \sum_{i=1}^n x^i y^i$
 $\leq |x|^2 + |y|^2 + 2|x| \cdot |y| \quad \text{by (2)}$
 $= (|x| + |y|)^2.$
- (4) $|ax| = \sqrt{\sum_{i=1}^n (ax^i)^2} = \sqrt{a^2 \sum_{i=1}^n (x^i)^2} = |a| \cdot |x|. \quad \blacksquare$

The quantity $\sum_{i=1}^n x^i y^i$ which appears in (2) is called the **inner product** of x and y and denoted $\langle x, y \rangle$. The most important properties of the inner product are the following.

1-2 Theorem. If x, x_1, x_2 and y, y_1, y_2 are vectors in \mathbf{R}^n and $a \in \mathbf{R}$, then

- (1) $\langle x, y \rangle = \langle y, x \rangle$ (symmetry).

- (2) $\langle ax, y \rangle = \langle x, ay \rangle = a\langle x, y \rangle$ (bilinearity).
 $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$
 $\langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle$
- (3) $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$ (positive definiteness).
- (4) $|x| = \sqrt{\langle x, x \rangle}$.
- (5) $\langle x, y \rangle = \frac{|x + y|^2 - |x - y|^2}{4}$ (polarization identity).

Proof

- (1) $\langle x, y \rangle = \sum_{i=1}^n x^i y^i = \sum_{i=1}^n y^i x^i = \langle y, x \rangle$.
 (2) By (1) it suffices to prove

$$\begin{aligned}\langle ax, y \rangle &= a\langle x, y \rangle, \\ \langle x_1 + x_2, y \rangle &= \langle x_1, y \rangle + \langle x_2, y \rangle.\end{aligned}$$

These follow from the equations

$$\begin{aligned}\langle ax, y \rangle &= \sum_{i=1}^n (ax^i) y^i = a \sum_{i=1}^n x^i y^i = a\langle x, y \rangle, \\ \langle x_1 + x_2, y \rangle &= \sum_{i=1}^n (x_1^i + x_2^i) y^i = \sum_{i=1}^n x_1^i y^i + \sum_{i=1}^n x_2^i y^i \\ &= \langle x_1, y \rangle + \langle x_2, y \rangle.\end{aligned}$$

(3) and (4) are left to the reader.

$$\begin{aligned}(5) \quad & \frac{|x + y|^2 - |x - y|^2}{4} \\ &= \frac{1}{4}[\langle x + y, x + y \rangle - \langle x - y, x - y \rangle] \quad \text{by (4)} \\ &= \frac{1}{4}[\langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle - (\langle x, x \rangle - 2\langle x, y \rangle + \langle y, y \rangle)] \\ &= \langle x, y \rangle. \quad \blacksquare\end{aligned}$$

We conclude this section with some important remarks about notation. The vector $(0, \dots, 0)$ will usually be denoted simply 0 . The **usual basis** of \mathbf{R}^n is e_1, \dots, e_n , where $e_i = (0, \dots, 1, \dots, 0)$, with the 1 in the i th place. If $T: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is a linear transformation, the matrix of T with respect to the usual bases of \mathbf{R}^n and \mathbf{R}^m is the $m \times n$ matrix $A = (a_{ij})$, where $T(e_i) = \sum_{j=1}^m a_{ji} e_j$ —the coefficients of $T(e_i)$

appear in the i th *column* of the matrix. If $S: \mathbf{R}^m \rightarrow \mathbf{R}^p$ has the $p \times m$ matrix B , then $S \circ T$ has the $p \times n$ matrix BA [here $S \circ T(x) = S(T(x))$; most books on linear algebra denote $S \circ T$ simply ST]. To find $T(x)$ one computes the $m \times 1$ matrix

$$\begin{pmatrix} y^1 \\ \vdots \\ y^m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix};$$

then $T(x) = (y^1, \dots, y^m)$. One notational convention greatly simplifies many formulas: if $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$, then (x, y) denotes

$$(x^1, \dots, x^n, y^1, \dots, y^m) \in \mathbf{R}^{n+m}.$$

Problems. 1-1.* Prove that $|x| \leq \sum_{i=1}^n |x^i|$.

1-2. When does equality hold in Theorem 1-1(3)? *Hint:* Re-examine the proof; the answer is not “when x and y are linearly dependent.”

1-3. Prove that $|x - y| \leq |x| + |y|$. When does equality hold?

1-4. Prove that $||x| - |y|| \leq |x - y|$.

1-5. The quantity $|y - x|$ is called the **distance** between x and y . Prove and interpret geometrically the “triangle inequality”: $|z - x| \leq |z - y| + |y - x|$.

1-6. Let f and g be integrable on $[a, b]$.

(a) Prove that $|\int_a^b f \cdot g| \leq (\int_a^b f^2)^{\frac{1}{2}} \cdot (\int_a^b g^2)^{\frac{1}{2}}$. *Hint:* Consider separately the cases $0 = \int_a^b (f - \lambda g)^2$ for some $\lambda \in \mathbf{R}$ and $0 < \int_a^b (f - \lambda g)^2$ for all $\lambda \in \mathbf{R}$.

(b) If equality holds, must $f = \lambda g$ for some $\lambda \in \mathbf{R}$? What if f and g are continuous?

(c) Show that Theorem 1-1(2) is a special case of (a).

1-7. A linear transformation $T: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is **norm preserving** if $|T(x)| = |x|$, and **inner product preserving** if $\langle Tx, Ty \rangle = \langle x, y \rangle$.

(a) Prove that T is norm preserving if and only if T is inner-product preserving.

(b) Prove that such a linear transformation T is 1-1 and T^{-1} is of the same sort.

1-8. If $x, y \in \mathbf{R}^n$ are non-zero, the **angle** between x and y , denoted $\angle(x, y)$, is defined as $\arccos(\langle x, y \rangle / (|x| \cdot |y|))$, which makes sense by Theorem 1-1(2). The linear transformation T is **angle preserving** if T is 1-1, and for $x, y \neq 0$ we have $\angle(Tx, Ty) = \angle(x, y)$.

(a) Prove that if T is norm preserving, then T is angle preserving.

(b) If there is a basis x_1, \dots, x_n of \mathbf{R}^n and numbers $\lambda_1, \dots, \lambda_n$ such that $Tx_i = \lambda_i x_i$, prove that T is angle preserving if and only if all $|\lambda_i|$ are equal.

(c) What are all angle preserving $T: \mathbf{R}^n \rightarrow \mathbf{R}^n$?

1-9. If $0 \leq \theta < \pi$, let $T: \mathbf{R}^2 \rightarrow \mathbf{R}^2$ have the matrix $\begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$.

Show that T is angle preserving and if $x \neq 0$, then $\angle(x, Tx) = \theta$.

1-10.* If $T: \mathbf{R}^m \rightarrow \mathbf{R}^n$ is a linear transformation, show that there is a number M such that $|T(h)| \leq M|h|$ for $h \in \mathbf{R}^m$. *Hint:* Estimate $|T(h)|$ in terms of $|h|$ and the entries in the matrix of T .

1-11. If $x, y \in \mathbf{R}^n$ and $z, w \in \mathbf{R}^m$, show that $\langle (x, z), (y, w) \rangle = \langle x, y \rangle + \langle z, w \rangle$ and $|(x, z)| = \sqrt{|x|^2 + |z|^2}$. Note that (x, z) and (y, w) denote points in \mathbf{R}^{n+m} .

1-12.* Let $(\mathbf{R}^n)^*$ denote the dual space of the vector space \mathbf{R}^n . If $x \in \mathbf{R}^n$, define $\varphi_x \in (\mathbf{R}^n)^*$ by $\varphi_x(y) = \langle x, y \rangle$. Define $T: \mathbf{R}^n \rightarrow (\mathbf{R}^n)^*$ by $T(x) = \varphi_x$. Show that T is a 1-1 linear transformation and conclude that every $\varphi \in (\mathbf{R}^n)^*$ is φ_x for a unique $x \in \mathbf{R}^n$.

1-13.* If $x, y \in \mathbf{R}^n$, then x and y are called **perpendicular** (or **orthogonal**) if $\langle x, y \rangle = 0$. If x and y are perpendicular, prove that $|x + y|^2 = |x|^2 + |y|^2$.

SUBSETS OF EUCLIDEAN SPACE

The closed interval $[a, b]$ has a natural analogue in \mathbf{R}^2 . This is the **closed rectangle** $[a, b] \times [c, d]$, defined as the collection of all pairs (x, y) with $x \in [a, b]$ and $y \in [c, d]$. More generally, if $A \subset \mathbf{R}^m$ and $B \subset \mathbf{R}^n$, then $A \times B \subset \mathbf{R}^{m+n}$ is defined as the set of all $(x, y) \in \mathbf{R}^{m+n}$ with $x \in A$ and $y \in B$. In particular, $\mathbf{R}^{m+n} = \mathbf{R}^m \times \mathbf{R}^n$. If $A \subset \mathbf{R}^m$, $B \subset \mathbf{R}^n$, and $C \subset \mathbf{R}^p$, then $(A \times B) \times C = A \times (B \times C)$, and both of these are denoted simply $A \times B \times C$; this convention is extended to the product of any number of sets. The set $[a_1, b_1] \times \dots \times [a_n, b_n] \subset \mathbf{R}^n$ is called a **closed rectangle** in \mathbf{R}^n , while the set $(a_1, b_1) \times \dots \times (a_n, b_n) \subset \mathbf{R}^n$ is called an **open rectangle**. More generally a set $U \subset \mathbf{R}^n$ is called **open** (Figure 1-1) if for each $x \in U$ there is an open rectangle A such that $x \in A \subset U$.

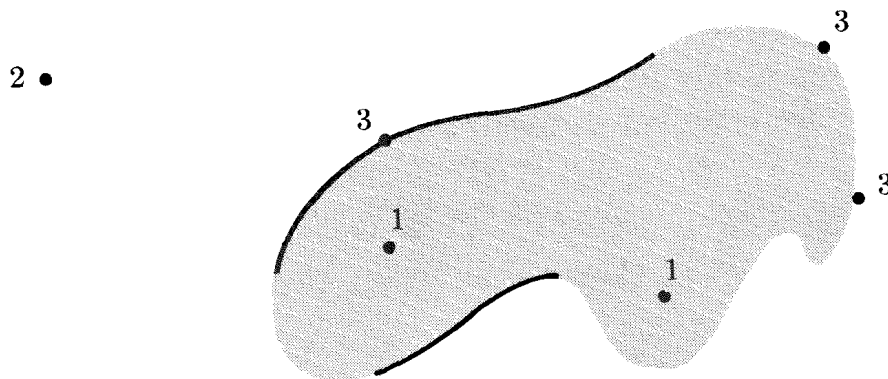
A subset C of \mathbf{R}^n is **closed** if $\mathbf{R}^n - C$ is open. For example, if C contains only finitely many points, then C is closed.

**FIGURE 1-1**

The reader should supply the proof that a closed rectangle in \mathbf{R}^n is indeed a closed set.

If $A \subset \mathbf{R}^n$ and $x \in \mathbf{R}^n$, then one of three possibilities must hold (Figure 1-2):

1. There is an open rectangle B such that $x \in B \subset A$.
2. There is an open rectangle B such that $x \in B \subset \mathbf{R}^n - A$.
3. If B is any open rectangle with $x \in B$, then B contains points of both A and $\mathbf{R}^n - A$.

**FIGURE 1-2**

Those points satisfying (1) constitute the **interior** of A , those satisfying (2) the **exterior** of A , and those satisfying (3) the **boundary** of A . Problems 1-16 to 1-18 show that these terms may sometimes have unexpected meanings.

It is not hard to see that the interior of any set A is open, and the same is true for the exterior of A , which is, in fact, the interior of $\mathbf{R}^n - A$. Thus (Problem 1-14) their union is open, and what remains, the boundary, must be closed.

A collection \mathcal{O} of open sets is an **open cover** of A (or, briefly, **covers** A) if every point $x \in A$ is in some open set in the collection \mathcal{O} . For example, if \mathcal{O} is the collection of all open intervals $(a, a + 1)$ for $a \in \mathbf{R}$, then \mathcal{O} is a cover of \mathbf{R} . Clearly no finite number of the open sets in \mathcal{O} will cover \mathbf{R} or, for that matter, any unbounded subset of \mathbf{R} . A similar situation can also occur for bounded sets. If \mathcal{O} is the collection of all open intervals $(1/n, 1 - 1/n)$ for all integers $n > 1$, then \mathcal{O} is an open cover of $(0,1)$, but again no finite collection of sets in \mathcal{O} will cover $(0,1)$. Although this phenomenon may not appear particularly scandalous, sets for which this state of affairs cannot occur are of such importance that they have received a special designation: a set A is called **compact** if every open cover \mathcal{O} contains a finite subcollection of open sets which also covers A .

A set with only finitely many points is obviously compact and so is the infinite set A which contains 0 and the numbers $1/n$ for all integers n (reason: if \mathcal{O} is a cover, then $0 \in U$ for some open set U in \mathcal{O} ; there are only finitely many other points of A not in U , each requiring at most one more open set).

Recognizing compact sets is greatly simplified by the following results, of which only the first has any depth (i.e., uses any facts about the real numbers).

1-3 Theorem (Heine-Borel). *The closed interval $[a,b]$ is compact.*

Proof. If \mathcal{O} is an open cover of $[a,b]$, let

$A = \{x: a \leq x \leq b \text{ and } [a,x] \text{ is covered by some finite number of open sets in } \mathcal{O}\}.$

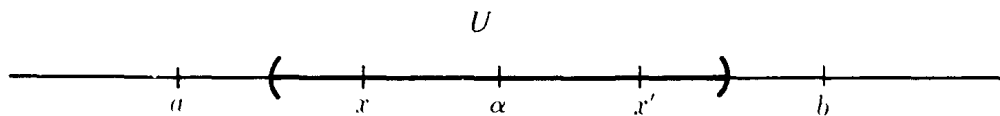


FIGURE 1-3

Note that $a \in A$ and that A is clearly bounded above (by b). We would like to show that $b \in A$. This is done by proving two things about $\alpha = \text{least upper bound of } A$; namely, (1) $\alpha \in A$ and (2) $b = \alpha$.

Since \mathcal{O} is a cover, $\alpha \in U$ for some U in \mathcal{O} . Then all points in some interval to the left of α are also in U (see Figure 1-3). Since α is the least upper bound of A , there is an x in this interval such that $x \in A$. Thus $[a, x]$ is covered by some finite number of open sets of \mathcal{O} , while $[x, \alpha]$ is covered by the single set U . Hence $[a, \alpha]$ is covered by a finite number of open sets of \mathcal{O} , and $\alpha \in A$. This proves (1).

To prove that (2) is true, suppose instead that $\alpha < b$. Then there is a point x' between α and b such that $[\alpha, x'] \subset U$. Since $\alpha \in A$, the interval $[a, \alpha]$ is covered by finitely many open sets of \mathcal{O} , while $[\alpha, x']$ is covered by U . Hence $x' \in A$, contradicting the fact that α is an upper bound of A . ■

If $B \subset \mathbf{R}^m$ is compact and $x \in \mathbf{R}^n$, it is easy to see that $\{x\} \times B \subset \mathbf{R}^{n+m}$ is compact. However, a much stronger assertion can be made.

1-4 Theorem. *If B is compact and \mathcal{O} is an open cover of $\{x\} \times B$, then there is an open set $U \subset \mathbf{R}^n$ containing x such that $U \times B$ is covered by a finite number of sets in \mathcal{O} .*

Proof. Since $\{x\} \times B$ is compact, we can assume at the outset that \mathcal{O} is finite, and we need only find the open set U such that $U \times B$ is covered by \mathcal{O} .

For each $y \in B$ the point (x, y) is in some open set W in \mathcal{O} . Since W is open, we have $(x, y) \in U_y \times V_y \subset W$ for some open rectangle $U_y \times V_y$. The sets V_y cover the compact set B , so a finite number V_{y_1}, \dots, V_{y_k} also cover B . Let $U = U_{y_1} \cap \dots \cap U_{y_k}$. Then if $(x', y') \in U \times B$, we have

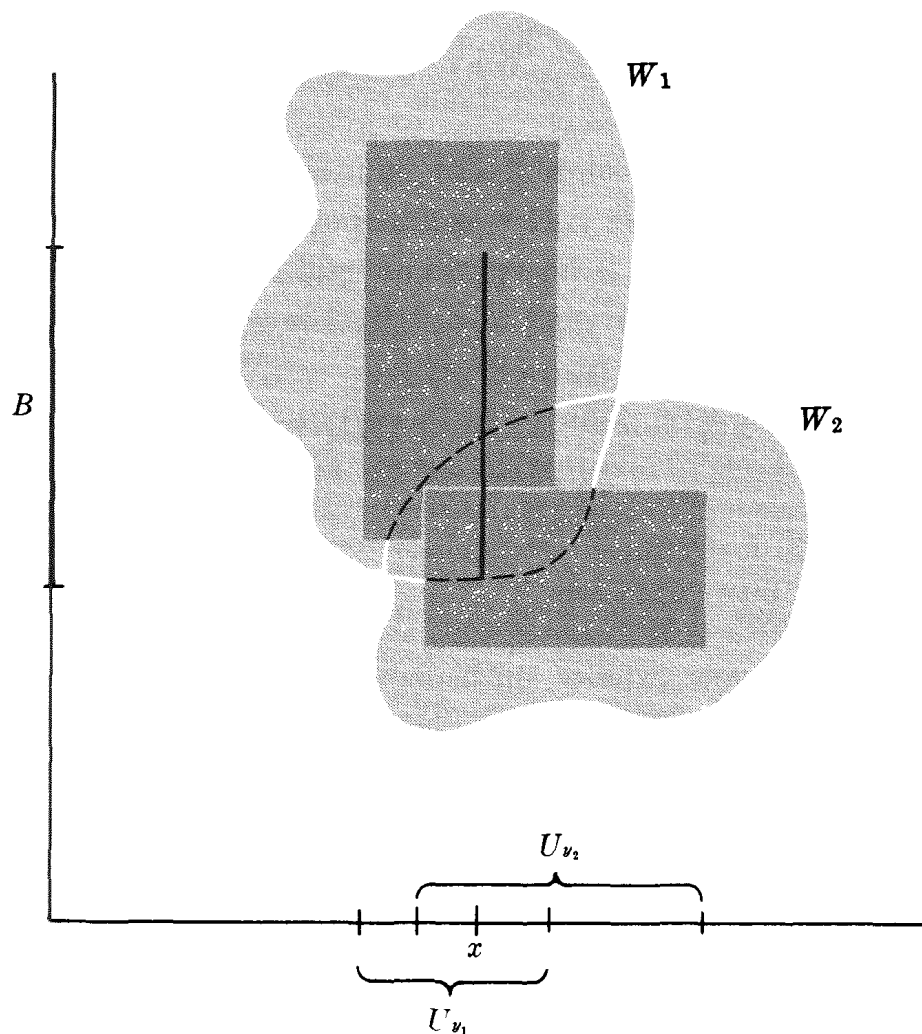


FIGURE 1-4

$y' \in V_{y_i}$ for some i (Figure 1-4), and certainly $x' \in U_{y_i}$. Hence $(x', y') \in U_{y_i} \times V_{y_i}$, which is contained in some W in \mathcal{O} . ■

1-5 Corollary. If $A \subset \mathbf{R}^n$ and $B \subset \mathbf{R}^m$ are compact, then $A \times B \subset \mathbf{R}^{n+m}$ is compact.

Proof. If \mathcal{O} is an open cover of $A \times B$, then \mathcal{O} covers $\{x\} \times B$ for each $x \in A$. By Theorem 1-4 there is an open set U_x containing x such that $U_x \times B$ is covered by finitely many sets in \mathcal{O} . Since A is compact, a finite number U_{x_1}, \dots, U_{x_n} of the U_x cover A . Since finitely many sets in \mathcal{O} cover each $U_{x_i} \times B$, finitely many cover all of $A \times B$. ■

1-6 Corollary. $A_1 \times \dots \times A_k$ is compact if each A_i is. In particular, a closed rectangle in \mathbf{R}^k is compact.

1-7 Corollary. *A closed bounded subset of \mathbf{R}^n is compact.*
(The converse is also true (Problem 1-20).)

Proof. If $A \subset \mathbf{R}^n$ is closed and bounded, then $A \subset B$ for some closed rectangle B . If \mathcal{O} is an open cover of A , then \mathcal{O} together with $\mathbf{R}^n - A$ is an open cover of B . Hence a finite number U_1, \dots, U_n of sets in \mathcal{O} , together with $\mathbf{R}^n - A$ perhaps, cover B . Then U_1, \dots, U_n cover A . ■

Problems. 1-14.* Prove that the union of any (even infinite) number of open sets is open. Prove that the intersection of two (and hence of finitely many) open sets is open. Give a counterexample for infinitely many open sets.

1-15. Prove that $\{x \in \mathbf{R}^n: |x - a| < r\}$ is open (see also Problem 1-27).

1-16. Find the interior, exterior, and boundary of the sets

$$\begin{aligned} &\{x \in \mathbf{R}^n: |x| \leq 1\} \\ &\{x \in \mathbf{R}^n: |x| = 1\} \\ &\{x \in \mathbf{R}^n: \text{each } x^i \text{ is rational}\}. \end{aligned}$$

1-17. Construct a set $A \subset [0,1] \times [0,1]$ such that A contains at most one point on each horizontal and each vertical line but boundary $A = [0,1] \times [0,1]$. *Hint:* It suffices to ensure that A contains points in each quarter of the square $[0,1] \times [0,1]$ and also in each sixteenth, etc.

1-18. If $A \subset [0,1]$ is the union of open intervals (a_i, b_i) such that each rational number in $(0,1)$ is contained in some (a_i, b_i) , show that boundary $A = [0,1] - A$.

1-19.* If A is a closed set that contains every rational number $r \in [0,1]$, show that $[0,1] \subset A$.

1-20. Prove the converse of Corollary 1-7: A compact subset of \mathbf{R}^n is closed and bounded (see also Problem 1-28).

1-21.* (a) If A is closed and $x \notin A$, prove that there is a number $d > 0$ such that $|y - x| \geq d$ for all $y \in A$.

(b) If A is closed, B is compact, and $A \cap B = \emptyset$, prove that there is $d > 0$ such that $|y - x| \geq d$ for all $y \in A$ and $x \in B$. *Hint:* For each $b \in B$ find an open set U containing b such that this relation holds for $x \in U \cap B$.

(c) Give a counterexample in \mathbf{R}^2 if A and B are closed but neither is compact.

1-22.* If U is open and $C \subset U$ is compact, show that there is a compact set D such that $C \subset \text{interior } D$ and $D \subset U$.

FUNCTIONS AND CONTINUITY

A **function** from \mathbf{R}^n to \mathbf{R}^m (sometimes called a (vector-valued) function of n variables) is a rule which associates to each point in \mathbf{R}^n some point in \mathbf{R}^m ; the point a function f associates to x is denoted $f(x)$. We write $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ (read “ f takes \mathbf{R}^n into \mathbf{R}^m ” or “ f , taking \mathbf{R}^n into \mathbf{R}^m ,” depending on context) to indicate that $f(x) \in \mathbf{R}^m$ is defined for $x \in \mathbf{R}^n$. The notation $f: A \rightarrow \mathbf{R}^m$ indicates that $f(x)$ is defined only for x in the set A , which is called the **domain** of f . If $B \subset A$, we define $f(B)$ as the set of all $f(x)$ for $x \in B$, and if $C \subset \mathbf{R}^m$ we define $f^{-1}(C) = \{x \in A: f(x) \in C\}$. The notation $f: A \rightarrow B$ indicates that $f(A) \subset B$.

A convenient representation of a function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ may be obtained by drawing a picture of its graph, the set of all 3-tuples of the form $(x, y, f(x, y))$, which is actually a figure in 3-space (see, e.g., Figures 2-1 and 2-2 of Chapter 2).

If $f, g: \mathbf{R}^n \rightarrow \mathbf{R}$, the functions $f + g$, $f - g$, $f \cdot g$, and f/g are defined precisely as in the one-variable case. If $f: A \rightarrow \mathbf{R}^m$ and $g: B \rightarrow \mathbf{R}^p$, where $B \subset \mathbf{R}^m$, then the **composition** $g \circ f$ is defined by $g \circ f(x) = g(f(x))$; the domain of $g \circ f$ is $A \cap f^{-1}(B)$. If $f: A \rightarrow \mathbf{R}^m$ is 1-1, that is, if $f(x) \neq f(y)$ when $x \neq y$, we define $f^{-1}: f(A) \rightarrow \mathbf{R}^n$ by the requirement that $f^{-1}(z)$ is the unique $x \in A$ with $f(x) = z$.

A function $f: A \rightarrow \mathbf{R}^m$ determines m **component functions** $f^1, \dots, f^m: A \rightarrow \mathbf{R}$ by $f(x) = (f^1(x), \dots, f^m(x))$. If conversely, m functions $g_1, \dots, g_m: A \rightarrow \mathbf{R}$ are given, there is a unique function $f: A \rightarrow \mathbf{R}^m$ such that $f^i = g_i$, namely $f(x) = (g_1(x), \dots, g_m(x))$. This function f will be denoted (g_1, \dots, g_m) , so that we always have $f = (f^1, \dots, f^m)$. If $\pi: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is the identity function, $\pi(x) = x$, then $\pi^i(x) = x^i$; the function π^i is called the i th **projection function**.

The notation $\lim_{x \rightarrow a} f(x) = b$ means, as in the one-variable case, that we can get $f(x)$ as close to b as desired, by choosing x sufficiently close to, but not equal to, a . In mathematical terms this means that for every number $\epsilon > 0$ there is a number $\delta > 0$ such that $|f(x) - b| < \epsilon$ for all x in the domain of f which

satisfy $0 < |x - a| < \delta$. A function $f: A \rightarrow \mathbf{R}^m$ is called **continuous** at $a \in A$ if $\lim_{x \rightarrow a} f(x) = f(a)$, and f is simply called con-

tinuous if it is continuous at each $a \in A$. One of the pleasant surprises about the concept of continuity is that it can be defined without using limits. It follows from the next theorem that $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is continuous if and only if $f^{-1}(U)$ is open whenever $U \subset \mathbf{R}^m$ is open; if the domain of f is not all of \mathbf{R}^n , a slightly more complicated condition is needed.

1-8 Theorem. *If $A \subset \mathbf{R}^n$, a function $f: A \rightarrow \mathbf{R}^m$ is continuous if and only if for every open set $U \subset \mathbf{R}^m$ there is some open set $V \subset \mathbf{R}^n$ such that $f^{-1}(U) = V \cap A$.*

Proof. Suppose f is continuous. If $a \in f^{-1}(U)$, then $f(a) \in U$. Since U is open, there is an open rectangle B with $f(a) \in B \subset U$. Since f is continuous at a , we can ensure that $f(x) \in B$, provided we choose x in some sufficiently small rectangle C containing a . Do this for each $a \in f^{-1}(U)$ and let V be the union of all such C . Clearly $f^{-1}(U) = V \cap A$. The converse is similar and is left to the reader. ■

The following consequence of Theorem 1-8 is of great importance.

1-9 Theorem. *If $f: A \rightarrow \mathbf{R}^m$ is continuous, where $A \subset \mathbf{R}^n$, and A is compact, then $f(A) \subset \mathbf{R}^m$ is compact.*

Proof. Let \mathcal{O} be an open cover of $f(A)$. For each open set U in \mathcal{O} there is an open set V_U such that $f^{-1}(U) = V_U \cap A$. The collection of all V_U is an open cover of A . Since A is compact, a finite number V_{U_1}, \dots, V_{U_n} cover A . Then U_1, \dots, U_n cover $f(A)$. ■

If $f: A \rightarrow \mathbf{R}$ is bounded, the extent to which f fails to be continuous at $a \in A$ can be measured in a precise way. For $\delta > 0$ let

$$M(a, f, \delta) = \sup\{f(x) : x \in A \text{ and } |x - a| < \delta\},$$

$$m(a, f, \delta) = \inf\{f(x) : x \in A \text{ and } |x - a| < \delta\}.$$

The **oscillation** $o(f,a)$ of f at a is defined by $o(f,a) = \lim_{\delta \rightarrow 0} [M(a,f,\delta) - m(a,f,\delta)]$. This limit always exists, since $M(a,f,\delta) - m(a,f,\delta)$ decreases as δ decreases. There are two important facts about $o(f,a)$.

1-10 Theorem. *The bounded function f is continuous at a if and only if $o(f,a) = 0$.*

Proof. Let f be continuous at a . For every number $\varepsilon > 0$ we can choose a number $\delta > 0$ so that $|f(x) - f(a)| < \varepsilon$ for all $x \in A$ with $|x - a| < \delta$; thus $M(a,f,\delta) - m(a,f,\delta) \leq 2\varepsilon$. Since this is true for every ε , we have $o(f,a) = 0$. The converse is similar and is left to the reader. ■

1-11 Theorem. *Let $A \subset \mathbf{R}^n$ be closed. If $f: A \rightarrow \mathbf{R}$ is any bounded function, and $\varepsilon > 0$, then $\{x \in A: o(f,x) \geq \varepsilon\}$ is closed.*

Proof. Let $B = \{x \in A: o(f,x) \geq \varepsilon\}$. We wish to show that $\mathbf{R}^n - B$ is open. If $x \in \mathbf{R}^n - B$, then either $x \notin A$ or else $x \in A$ and $o(f,x) < \varepsilon$. In the first case, since A is closed, there is an open rectangle C containing x such that $C \subset \mathbf{R}^n - A \subset \mathbf{R}^n - B$. In the second case there is a $\delta > 0$ such that $M(x,f,\delta) - m(x,f,\delta) < \varepsilon$. Let C be an open rectangle containing x such that $|x - y| < \delta$ for all $y \in C$. Then if $y \in C$ there is a δ_1 such that $|x - z| < \delta$ for all z satisfying $|z - y| < \delta_1$. Thus $M(y,f,\delta_1) - m(y,f,\delta_1) < \varepsilon$, and consequently $o(y,f) < \varepsilon$. Therefore $C \subset \mathbf{R}^n - B$. ■

Problems. 1-23. If $f: A \rightarrow \mathbf{R}^m$ and $a \in A$, show that $\lim_{x \rightarrow a} f(x) = b$ if and only if $\lim_{x \rightarrow a} f^i(x) = b^i$ for $i = 1, \dots, m$.

1-24. Prove that $f: A \rightarrow \mathbf{R}^m$ is continuous at a if and only if each f^i is.

1-25. Prove that a linear transformation $T: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is continuous.

Hint: Use Problem 1-10.

1-26. Let $A = \{(x,y) \in \mathbf{R}^2: x > 0 \text{ and } 0 < y < x^2\}$.

(a) Show that every straight line through $(0,0)$ contains an interval around $(0,0)$ which is in $\mathbf{R}^2 - A$.

(b) Define $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ by $f(x) = 0$ if $x \notin A$ and $f(x) = 1$ if $x \in A$. For $h \in \mathbf{R}^2$ define $g_h: \mathbf{R} \rightarrow \mathbf{R}$ by $g_h(t) = f(th)$. Show that each g_h is continuous at 0, but f is not continuous at $(0,0)$.

- 1-27. Prove that $\{x \in \mathbf{R}^n: |x - a| < r\}$ is open by considering the function $f: \mathbf{R}^n \rightarrow \mathbf{R}$ with $f(x) = |x - a|$.
- 1-28. If $A \subset \mathbf{R}^n$ is not closed, show that there is a continuous function $f: A \rightarrow \mathbf{R}$ which is unbounded. *Hint:* If $x \in \mathbf{R}^n - A$ but $x \notin \text{interior}(\mathbf{R}^n - A)$, let $f(y) = 1/|y - x|$.
- 1-29. If A is compact, prove that every continuous function $f: A \rightarrow \mathbf{R}$ takes on a maximum and a minimum value.
- 1-30. Let $f: [a, b] \rightarrow \mathbf{R}$ be an increasing function. If $x_1, \dots, x_n \in [a, b]$ are distinct, show that $\sum_{i=1}^n (f(x_i) - f(x_{i-1})) < f(b) - f(a)$.

2

Differentiation

BASIC DEFINITIONS

Recall that a function $f: \mathbf{R} \rightarrow \mathbf{R}$ is differentiable at $a \in \mathbf{R}$ if there is a number $f'(a)$ such that

$$(1) \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = f'(a).$$

This equation certainly makes no sense in the general case of a function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, but can be reformulated in a way that does. If $\lambda: \mathbf{R} \rightarrow \mathbf{R}$ is the linear transformation defined by $\lambda(h) = f'(a) \cdot h$, then equation (1) is equivalent to

$$(2) \lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - \lambda(h)}{h} = 0.$$

Equation (2) is often interpreted as saying that $\lambda + f(a)$ is a good approximation to f at a (see Problem 2-9). Henceforth we focus our attention on the linear transformation λ and reformulate the definition of differentiability as follows.

A function $f: \mathbf{R} \rightarrow \mathbf{R}$ is differentiable at $a \in \mathbf{R}$ if there is a linear transformation $\lambda: \mathbf{R} \rightarrow \mathbf{R}$ such that

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - \lambda(h)}{h} = 0.$$

In this form the definition has a simple generalization to higher dimensions:

A function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is **differentiable** at $a \in \mathbf{R}^n$ if there is a linear transformation $\lambda: \mathbf{R}^n \rightarrow \mathbf{R}^m$ such that

$$\lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - \lambda(h)|}{|h|} = 0.$$

Note that h is a point of \mathbf{R}^n and $f(a+h) - f(a) - \lambda(h)$ a point of \mathbf{R}^m , so the norm signs are essential. The linear transformation λ is denoted $Df(a)$ and called the **derivative** of f at a . The justification for the phrase “the linear transformation λ ” is

2-1 Theorem. *If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable at $a \in \mathbf{R}^n$ there is a unique linear transformation $\lambda: \mathbf{R}^n \rightarrow \mathbf{R}^m$ such that*

$$\lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - \lambda(h)|}{|h|} = 0.$$

Proof. Suppose $\mu: \mathbf{R}^n \rightarrow \mathbf{R}^m$ satisfies

$$\lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - \mu(h)|}{|h|} = 0.$$

If $d(h) = f(a+h) - f(a)$, then

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{|\lambda(h) - \mu(h)|}{|h|} &= \lim_{h \rightarrow 0} \frac{|\lambda(h) - d(h) + d(h) - \mu(h)|}{|h|} \\ &\leq \lim_{h \rightarrow 0} \frac{|\lambda(h) - d(h)|}{|h|} + \lim_{h \rightarrow 0} \frac{|d(h) - \mu(h)|}{|h|} \\ &= 0. \end{aligned}$$

If $x \in \mathbf{R}^n$, then $tx \rightarrow 0$ as $t \rightarrow 0$. Hence for $x \neq 0$ we have

$$0 = \lim_{t \rightarrow 0} \frac{|\lambda(tx) - \mu(tx)|}{|tx|} = \frac{|\lambda(x) - \mu(x)|}{|x|}.$$

Therefore $\lambda(x) = \mu(x)$. ■

We shall later discover a simple way of finding $Df(a)$. For the moment let us consider the function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ defined by $f(x, y) = \sin x$. Then $Df(a, b) = \lambda$ satisfies $\lambda(x, y) = (\cos a) \cdot x$. To prove this, note that

$$\begin{aligned} \lim_{(h,k) \rightarrow 0} \frac{|f(a+h, b+k) - f(a, b) - \lambda(h, k)|}{|(h, k)|} \\ = \lim_{(h,k) \rightarrow 0} \frac{|\sin(a+h) - \sin a - (\cos a) \cdot h|}{|(h, k)|}. \end{aligned}$$

Since $\sin'(a) = \cos a$, we have

$$\lim_{h \rightarrow 0} \frac{|\sin(a+h) - \sin a - (\cos a) \cdot h|}{|h|} = 0.$$

Since $|(h, k)| \geq |h|$, it is also true that

$$\lim_{h \rightarrow 0} \frac{|\sin(a+h) - \sin a - (\cos a) \cdot h|}{|(h, k)|} = 0.$$

It is often convenient to consider the matrix of $Df(a)$: $\mathbf{R}^n \rightarrow \mathbf{R}^m$ with respect to the usual bases of \mathbf{R}^n and \mathbf{R}^m . This $m \times n$ matrix is called the **Jacobian matrix** of f at a , and denoted $f'(a)$. If $f(x, y) = \sin x$, then $f'(a, b) = (\cos a, 0)$. If $f: \mathbf{R} \rightarrow \mathbf{R}$, then $f'(a)$ is a 1×1 matrix whose single entry is the number which is denoted $f'(a)$ in elementary calculus.

The definition of $Df(a)$ could be made if f were defined only in some open set containing a . Considering only functions defined on \mathbf{R}^n streamlines the statement of theorems and produces no real loss of generality. It is convenient to define a function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ to be differentiable on A if f is differentiable at a for each $a \in A$. If $f: A \rightarrow \mathbf{R}^m$, then f is called differentiable if f can be extended to a differentiable function on some open set containing A .

Problems. 2-1.* Prove that if $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable at $a \in \mathbf{R}^n$, then it is continuous at a . *Hint:* Use Problem 1-10.

2-2. A function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ is **independent of the second variable** if for each $x \in \mathbf{R}$ we have $f(x, y_1) = f(x, y_2)$ for all $y_1, y_2 \in \mathbf{R}$. Show that f is independent of the second variable if and only if there is a function $g: \mathbf{R} \rightarrow \mathbf{R}$ such that $f(x, y) = g(x)$. What is $f'(a, b)$ in terms of g' ?

- 2-3.** Define when a function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ is independent of the first variable and find $f'(a,b)$ for such f . Which functions are independent of the first variable and also of the second variable?
- 2-4.** Let g be a continuous real-valued function on the unit circle $\{x \in \mathbf{R}^2: |x| = 1\}$ such that $g(0,1) = g(1,0) = 0$ and $g(-x) = -g(x)$. Define $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ by

$$f(x) = \begin{cases} |x| \cdot g\left(\frac{x}{|x|}\right) & x \neq 0, \\ 0 & x = 0. \end{cases}$$

- (a) If $x \in \mathbf{R}^2$ and $h: \mathbf{R} \rightarrow \mathbf{R}$ is defined by $h(t) = f(tx)$, show that h is differentiable.
- (b) Show that f is not differentiable at $(0,0)$ unless $g = 0$.
Hint: First show that $Df(0,0)$ would have to be 0 by considering (h,k) with $k = 0$ and then with $h = 0$.
- 2-5.** Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by

$$f(x,y) = \begin{cases} \frac{x|y|}{\sqrt{x^2 + y^2}} & (x,y) \neq 0, \\ 0 & (x,y) = 0. \end{cases}$$

Show that f is a function of the kind considered in Problem 2-4, so that f is not differentiable at $(0,0)$.

- 2-6.** Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by $f(x,y) = \sqrt{|xy|}$. Show that f is not differentiable at $(0,0)$.
- 2-7.** Let $f: \mathbf{R}^n \rightarrow \mathbf{R}$ be a function such that $|f(x)| \leq |x|^2$. Show that f is differentiable at 0.
- 2-8.** Let $f: \mathbf{R} \rightarrow \mathbf{R}^2$. Prove that f is differentiable at $a \in \mathbf{R}$ if and only if f^1 and f^2 are, and that in this case

$$f'(a) = \begin{pmatrix} (f^1)'(a) \\ (f^2)'(a) \end{pmatrix}.$$

- 2-9.** Two functions $f, g: \mathbf{R} \rightarrow \mathbf{R}$ are **equal up to n th order** at a if

$$\lim_{h \rightarrow 0} \frac{f(a+h) - g(a+h)}{h^n} = 0.$$

(a) Show that f is differentiable at a if and only if there is a function g of the form $g(x) = a_0 + a_1(x-a)$ such that f and g are equal up to first order at a .

(b) If $f'(a), \dots, f^{(n)}(a)$ exist, show that f and the function g defined by

$$g(x) = \sum_{i=0}^n \frac{f^{(i)}(a)}{i!} (x-a)^i$$

are equal up to n th order at a . *Hint:* The limit

$$\lim_{x \rightarrow a} \frac{f(x) - \sum_{i=0}^{n-1} \frac{f^{(i)}(a)}{i!} (x-a)^i}{(x-a)^n}$$

may be evaluated by L'Hospital's rule.

BASIC THEOREMS

2-2 Theorem (Chain Rule). *If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable at a , and $g: \mathbf{R}^m \rightarrow \mathbf{R}^p$ is differentiable at $f(a)$, then the composition $g \circ f: \mathbf{R}^n \rightarrow \mathbf{R}^p$ is differentiable at a , and*

$$D(g \circ f)(a) = Dg(f(a)) \circ Df(a).$$

Remark. This equation can be written

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a).$$

If $m = n = p = 1$, we obtain the old chain rule.

Proof. Let $b = f(a)$, let $\lambda = Df(a)$, and let $\mu = Dg(f(a))$. If we define

- (1) $\varphi(x) = f(x) - f(a) - \lambda(x-a)$,
- (2) $\psi(y) = g(y) - g(b) - \mu(y-b)$,
- (3) $\rho(x) = g \circ f(x) - g \circ f(a) - \mu \circ \lambda(x-a)$,

then

$$(4) \lim_{x \rightarrow a} \frac{|\varphi(x)|}{|x-a|} = 0,$$

$$(5) \lim_{y \rightarrow b} \frac{|\psi(y)|}{|y-b|} = 0,$$

and we must show that

$$\lim_{x \rightarrow a} \frac{|\rho(x)|}{|x-a|} = 0.$$

Now

$$\begin{aligned} \rho(x) &= g(f(x)) - g(b) - \mu(\lambda(x-a)) \\ &= g(f(x)) - g(b) - \mu(f(x) - f(a) - \varphi(x)) \quad \text{by (1)} \\ &= [g(f(x)) - g(b) - \mu(f(x) - f(a))] + \mu(\varphi(x)) \\ &= \psi(f(x)) + \mu(\varphi(x)) \quad \text{by (2)}. \end{aligned}$$

Thus we must prove

$$(6) \lim_{x \rightarrow a} \frac{|\psi(f(x))|}{|x - a|} = 0,$$

$$(7) \lim_{x \rightarrow a} \frac{|\mu(\varphi(x))|}{|x - a|} = 0.$$

Equation (7) follows easily from (4) and Problem 1-10. If $\varepsilon > 0$ it follows from (5) that for some $\delta > 0$ we have

$$|\psi(f(x))| < \varepsilon |f(x) - b| \quad \text{if } |f(x) - b| < \delta,$$

which is true if $|x - a| < \delta_1$, for a suitable δ_1 . Then

$$\begin{aligned} |\psi(f(x))| &< \varepsilon |f(x) - b| \\ &= \varepsilon |\varphi(x) + \lambda(x - a)| \\ &\leq \varepsilon |\varphi(x)| + \varepsilon M |x - a| \end{aligned}$$

for some M , by Problem 1-10. Equation (6) now follows easily. ■

2-3 Theorem

(1) If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is a constant function (that is, if for some $y \in \mathbf{R}^m$ we have $f(x) = y$ for all $x \in \mathbf{R}^n$), then

$$Df(a) = 0.$$

(2) If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is a linear transformation, then

$$Df(a) = f.$$

(3) If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, then f is differentiable at $a \in \mathbf{R}^n$ if and only if each f^i is, and

$$Df(a) = (Df^1(a), \dots, Df^m(a)).$$

Thus $f'(a)$ is the $m \times n$ matrix whose i th row is $(f^i)'(a)$.

(4) If $s: \mathbf{R}^2 \rightarrow \mathbf{R}$ is defined by $s(x, y) = x + y$, then

$$Ds(a, b) = s.$$

(5) If $p: \mathbf{R}^2 \rightarrow \mathbf{R}$ is defined by $p(x, y) = x \cdot y$, then

$$Dp(a, b)(x, y) = bx + ay.$$

Thus $p'(a, b) = (b, a)$.

Proof

$$(1) \lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - 0|}{|h|} = \lim_{h \rightarrow 0} \frac{|y - y - 0|}{|h|} = 0.$$

$$(2) \lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - f(h)|}{|h|} \\ = \lim_{h \rightarrow 0} \frac{|f(a) + f(h) - f(a) - f(h)|}{|h|} = 0.$$

(3) If each f^i is differentiable at a and

$$\lambda = (Df^1(a), \dots, Df^m(a)),$$

then

$$\begin{aligned} f(a+h) - f(a) - \lambda(h) \\ = (f^1(a+h) - f^1(a) - Df^1(a)(h), \dots, \\ f^m(a+h) - f^m(a) - Df^m(a)(h)). \end{aligned}$$

Therefore

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{|f(a+h) - f(a) - \lambda(h)|}{|h|} \\ \leq \lim_{h \rightarrow 0} \sum_{i=1}^m \frac{|f^i(a+h) - f^i(a) - Df^i(a)(h)|}{|h|} = 0. \end{aligned}$$

If, on the other hand, f is differentiable at a , then $f^i = \pi^i \circ f$ is differentiable at a by (2) and Theorem 2-2.

(4) follows from (2).

(5) Let $\lambda(x,y) = bx + ay$. Then

$$\begin{aligned} \lim_{(h,k) \rightarrow 0} \frac{|p(a+h, b+k) - p(a,b) - \lambda(h,k)|}{|(h,k)|} \\ = \lim_{(h,k) \rightarrow 0} \frac{|hk|}{|(h,k)|}. \end{aligned}$$

Now

$$|hk| \leq \begin{cases} |h|^2 & \text{if } |k| \leq |h|, \\ |k|^2 & \text{if } |h| \leq |k|. \end{cases}$$

Hence $|hk| \leq |h|^2 + |k|^2$. Therefore

$$\frac{|hk|}{|(h,k)|} \leq \frac{h^2 + k^2}{\sqrt{h^2 + k^2}} = \sqrt{h^2 + k^2},$$

so

$$\lim_{(h,k) \rightarrow 0} \frac{|hk|}{|(h,k)|} = 0. \quad \blacksquare$$

2-4 Corollary. If $f, g: \mathbf{R}^n \rightarrow \mathbf{R}$ are differentiable at a , then

$$\begin{aligned} D(f + g)(a) &= Df(a) + Dg(a), \\ D(f \cdot g)(a) &= g(a)Df(a) + f(a)Dg(a). \end{aligned}$$

If, moreover, $g(a) \neq 0$, then

$$D(f/g)(a) = \frac{g(a)Df(a) - f(a)Dg(a)}{[g(a)]^2}.$$

Proof. We will prove the first equation and leave the others to the reader. Since $f + g = s \circ (f, g)$, we have

$$\begin{aligned} D(f + g)(a) &= Ds(f(a), g(a)) \circ D(f, g)(a) \\ &= s \circ (Df(a), Dg(a)) \\ &= Df(a) + Dg(a). \quad \blacksquare \end{aligned}$$

We are now assured of the differentiability of those functions $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, whose component functions are obtained by addition, multiplication, division, and composition, from the functions π^i (which are linear transformations) and the functions which we can already differentiate by elementary calculus. Finding $Df(x)$ or $f'(x)$, however, may be a fairly formidable task. For example, let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by $f(x, y) = \sin(xy^2)$. Since $f = \sin \circ (\pi^1 \cdot [\pi^2]^2)$, we have

$$\begin{aligned} f'(a, b) &= \sin'(ab^2) \cdot [b^2(\pi^1)'(a, b) + a([\pi^2]^2)'(a, b)] \\ &= \sin'(ab^2) \cdot [b^2(\pi^1)'(a, b) + 2ab(\pi^2)'(a, b)] \\ &= (\cos(ab^2)) \cdot [b^2(1, 0) + 2ab(0, 1)] \\ &= (b^2 \cos(ab^2), 2ab \cos(ab^2)). \end{aligned}$$

Fortunately, we will soon discover a much simpler method of computing f' .

Problems. 2-10. Use the theorems of this section to find f' for the following:

- (a) $f(x, y, z) = x^y$.
- (b) $f(x, y, z) = (x^y, z)$.
- (c) $f(x, y) = \sin(x \sin y)$.
- (d) $f(x, y, z) = \sin(x \sin(y \sin z))$.
- (e) $f(x, y, z) = x^{y^z}$.
- (f) $f(x, y, z) = x^{y+z}$.
- (g) $f(x, y, z) = (x + y)^z$.
- (h) $f(x, y) = \sin(xy)$.
- (i) $f(x, y) = [\sin(xy)]^{\cos z}$.
- (j) $f(x, y) = (\sin(xy), \sin(x \sin y), x^y)$.

2-11. Find f' for the following (where $g: \mathbf{R} \rightarrow \mathbf{R}$ is continuous):

- (a) $f(x, y) = \int_a^{x+y} g$.
- (b) $f(x, y) = \int_a^{x \cdot y} g$.
- (c) $f(x, y, z) = \int_{xy}^{\sin(x \sin(y \sin z))} g$.

2-12. A function $f: \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}^p$ is **bilinear** if for $x, x_1, x_2 \in \mathbf{R}^n$, $y, y_1, y_2 \in \mathbf{R}^m$, and $a \in \mathbf{R}$ we have

$$\begin{aligned} f(ax, y) &= af(x, y) = f(x, ay), \\ f(x_1 + x_2, y) &= f(x_1, y) + f(x_2, y), \\ f(x, y_1 + y_2) &= f(x, y_1) + f(x, y_2). \end{aligned}$$

(a) Prove that if f is bilinear, then

$$\lim_{(h, k) \rightarrow 0} \frac{|f(h, k)|}{|(h, k)|} = 0.$$

(b) Prove that $Df(a, b)(x, y) = f(a, y) + f(x, b)$.

(c) Show that the formula for $Dp(a, b)$ in Theorem 2-3 is a special case of (b).

2-13. Define $IP: \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}$ by $IP(x, y) = \langle x, y \rangle$.

- (a) Find $D(IP)(a, b)$ and $(IP)'(a, b)$.
- (b) If $f, g: \mathbf{R} \rightarrow \mathbf{R}^n$ are differentiable and $h: \mathbf{R} \rightarrow \mathbf{R}$ is defined by $h(t) = \langle f(t), g(t) \rangle$, show that

$$h'(a) = \langle f'(a)^T, g(a) \rangle + \langle f(a), g'(a)^T \rangle.$$

(Note that $f'(a)$ is an $n \times 1$ matrix; its transpose $f'(a)^T$ is a $1 \times n$ matrix, which we consider as a member of \mathbf{R}^n .)

(c) If $f: \mathbf{R} \rightarrow \mathbf{R}^n$ is differentiable and $|f(t)| = 1$ for all t , show that $\langle f'(t)^T, f(t) \rangle = 0$.

(d) Exhibit a differentiable function $f: \mathbf{R} \rightarrow \mathbf{R}$ such that the function $|f|$ defined by $|f|(t) = |f(t)|$ is not differentiable.

2-14. Let E_i , $i = 1, \dots, k$ be Euclidean spaces of various dimensions. A function $f: E_1 \times \dots \times E_k \rightarrow \mathbf{R}^p$ is called **multilinear** if for each choice of $x_j \in E_j$, $j \neq i$ the function $g: E_i \rightarrow \mathbf{R}^p$ defined by $g(x) = f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_k)$ is a linear transformation.

(a) If f is multilinear and $i \neq j$, show that for $h = (h_1, \dots, h_k)$, with $h_l \in E_l$, we have

$$\lim_{h \rightarrow 0} \frac{|f(a_1, \dots, h_i, \dots, h_j, \dots, a_k)|}{|h|} = 0.$$

Hint: If $g(x, y) = f(a_1, \dots, x, \dots, y, \dots, a_k)$, then g is bilinear.

(b) Prove that

$$Df(a_1, \dots, a_k)(x_1, \dots, x_k) = \sum_{i=1}^k f(a_1, \dots, a_{i-1}, x_i, a_{i+1}, \dots, a_k).$$

2-15. Regard an $n \times n$ matrix as a point in the n -fold product $\mathbf{R}^n \times \dots \times \mathbf{R}^n$ by considering each row as a member of \mathbf{R}^n .

(a) Prove that $\det: \mathbf{R}^n \times \dots \times \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable and

$$D(\det)(a_1, \dots, a_n)(x_1, \dots, x_n) = \sum_{i=1}^n \det \begin{pmatrix} a_1 \\ \vdots \\ x_i \\ \vdots \\ a_n \end{pmatrix}.$$

(b) If $a_{ij}: \mathbf{R} \rightarrow \mathbf{R}$ are differentiable and $f(t) = \det(a_{ij}(t))$, show that

$$f'(t) = \sum_{j=1}^n \det \begin{pmatrix} a_{11}(t), \dots, a_{1n}(t) \\ \vdots \\ a_{j1}'(t), \dots, a_{jn}'(t) \\ \vdots \\ a_{n1}(t), \dots, a_{nn}(t) \end{pmatrix}.$$

(c) If $\det(a_{ij}(t)) \neq 0$ for all t and $b_1, \dots, b_n: \mathbf{R} \rightarrow \mathbf{R}$ are differentiable, let $s_1, \dots, s_n: \mathbf{R} \rightarrow \mathbf{R}$ be the functions such that $s_1(t), \dots, s_n(t)$ are the solutions of the equations

$$\sum_{j=1}^n a_{ji}(t)s_j(t) = b_i(t) \quad i = 1, \dots, n.$$

Show that s_i is differentiable and find $s_i'(t)$.

- 2-16.** Suppose $f: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is differentiable and has a differentiable inverse $f^{-1}: \mathbf{R}^n \rightarrow \mathbf{R}^n$. Show that $(f^{-1})'(a) = [f'(f^{-1}(a))]^{-1}$.
Hint: $f \circ f^{-1}(x) = x$.

PARTIAL DERIVATIVES

We begin the attack on the problem of finding derivatives “one variable at a time.” If $f: \mathbf{R}^n \rightarrow \mathbf{R}$ and $a \in \mathbf{R}^n$, the limit

$$\lim_{h \rightarrow 0} \frac{f(a^1, \dots, a^i + h, \dots, a^n) - f(a^1, \dots, a^n)}{h},$$

if it exists, is denoted $D_i f(a)$, and called the i th **partial derivative** of f at a . It is important to note that $D_i f(a)$ is the ordinary derivative of a certain function; in fact, if $g(x) = f(a^1, \dots, x, \dots, a^n)$, then $D_i f(a) = g'(a^i)$. This means that $D_i f(a)$ is the slope of the tangent line at $(a, f(a))$ to the curve obtained by intersecting the graph of f with the plane $x^j = a^j, j \neq i$ (Figure 2-1). It also means that computation of $D_i f(a)$ is a problem we can already solve. If $f(x^1, \dots, x^n)$ is

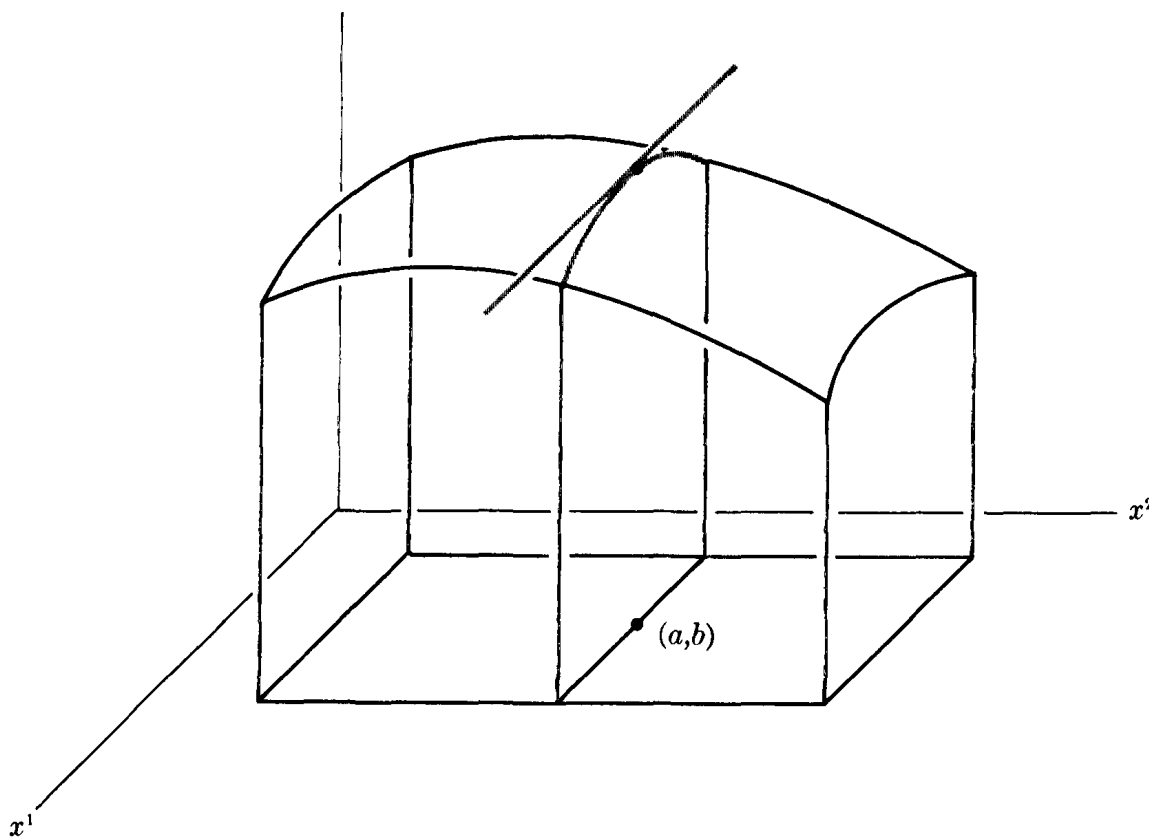


FIGURE 2-1

given by some formula involving x^1, \dots, x^n , then we find $D_i f(x^1, \dots, x^n)$ by differentiating the function whose value at x^i is given by the formula when all x^j , for $j \neq i$, are thought of as constants. For example, if $f(x, y) = \sin(xy^2)$, then $D_1 f(x, y) = y^2 \cos(xy^2)$ and $D_2 f(x, y) = 2xy \cos(xy^2)$. If, instead, $f(x, y) = x^y$, then $D_1 f(x, y) = yx^{y-1}$ and $D_2 f(x, y) = x^y \log x$.

With a little practice (e.g., the problems at the end of this section) you should acquire as great a facility for computing $D_i f$ as you already have for computing ordinary derivatives.

If $D_i f(x)$ exists for all $x \in \mathbf{R}^n$, we obtain a function $D_i f: \mathbf{R}^n \rightarrow \mathbf{R}$. The j th partial derivative of this function at x , that is, $D_j(D_i f)(x)$, is often denoted $D_{i,j} f(x)$. Note that this notation reverses the order of i and j . As a matter of fact, the order is usually irrelevant, since most functions (an exception is given in the problems) satisfy $D_{i,j} f = D_{j,i} f$. There are various delicate theorems ensuring this equality; the following theorem is quite adequate. We state it here but postpone the proof until later (Problem 3-28).

2-5 Theorem. *If $D_{i,j} f$ and $D_{j,i} f$ are continuous in an open set containing a , then*

$$D_{i,j} f(a) = D_{j,i} f(a).$$

The function $D_{i,j} f$ is called a **second-order (mixed) partial derivative** of f . Higher-order (mixed) partial derivatives are defined in the obvious way. Clearly Theorem 2-5 can be used to prove the equality of higher-order mixed partial derivatives under appropriate conditions. The order of i_1, \dots, i_k is completely immaterial in $D_{i_1, \dots, i_k} f$ if f has continuous partial derivatives of all orders. A function with this property is called a C^∞ function. In later chapters it will frequently be convenient to restrict our attention to C^∞ functions.

Partial derivatives will be used in the next section to find derivatives. They also have another important use—finding maxima and minima of functions.

2-6 Theorem. Let $A \subset \mathbf{R}^n$. If the maximum (or minimum) of $f: A \rightarrow \mathbf{R}$ occurs at a point a in the interior of A and $D_i f(a)$ exists, then $D_i f(a) = 0$.

Proof. Let $g_i(x) = f(a^1, \dots, x, \dots, a^n)$. Clearly g_i has a maximum (or minimum) at a^i , and g_i is defined in an open interval containing a^i . Hence $0 = g_i'(a^i) = D_i f(a)$. ■

The reader is reminded that the converse of Theorem 2-6 is false even if $n = 1$ (if $f: \mathbf{R} \rightarrow \mathbf{R}$ is defined by $f(x) = x^3$, then $f'(0) = 0$, but 0 is not even a local maximum or minimum). If $n > 1$, the converse of Theorem 2-6 may fail to be true in a rather spectacular way. Suppose, for example, that $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ is defined by $f(x, y) = x^2 - y^2$ (Figure 2-2). Then $D_1 f(0, 0) = 0$ because g_1 has a minimum at 0, while $D_2 f(0, 0) = 0$ because g_2 has a maximum at 0. Clearly $(0, 0)$ is neither a relative maximum nor a relative minimum.

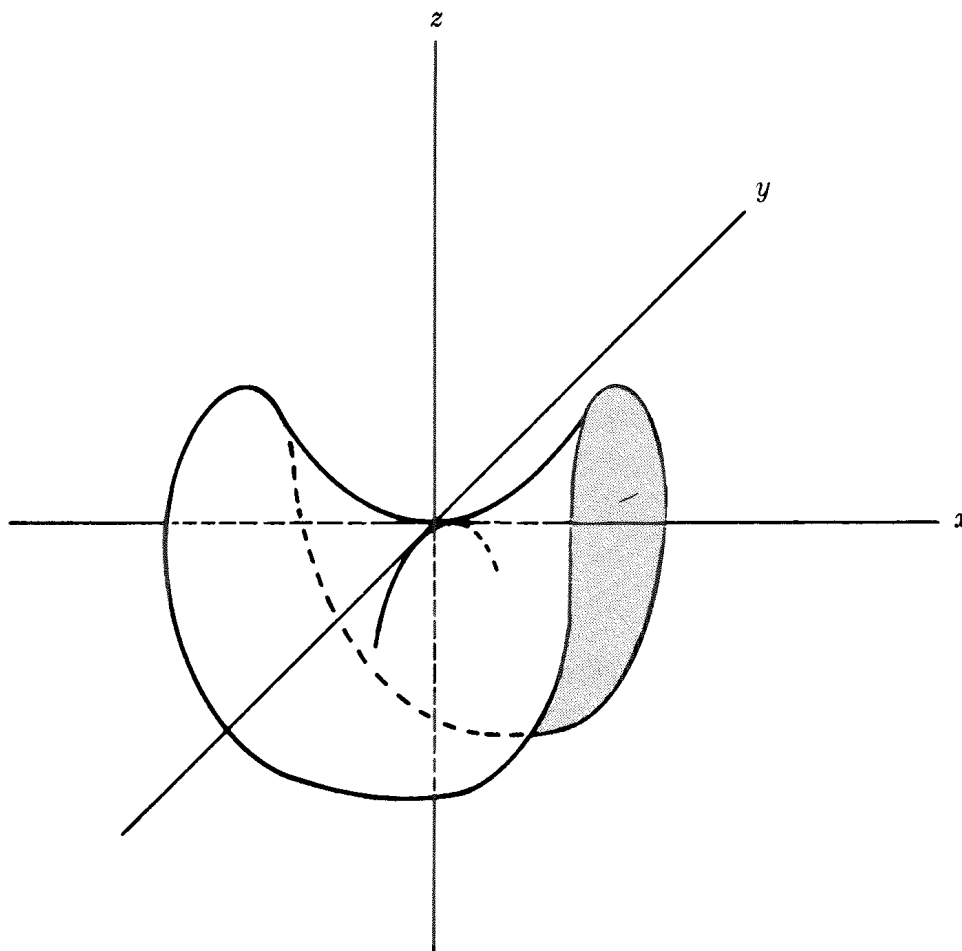


FIGURE 2-2

If Theorem 2-6 is used to find the maximum or minimum of f on A , the values of f at boundary points must be examined separately—a formidable task, since the boundary of A may be all of A ! Problem 2-27 indicates one way of doing this, and Problem 5-16 states a superior method which can often be used.

Problems. 2-17. Find the partial derivatives of the following functions:

- (a) $f(x, y, z) = x^y$.
- (b) $f(x, y, z) = z$.
- (c) $f(x, y) = \sin(x \sin y)$.
- (d) $f(x, y, z) = \sin(x \sin(y \sin z))$.
- (e) $f(x, y, z) = x^{y^z}$.
- (f) $f(x, y, z) = x^{y+z}$.
- (g) $f(x, y, z) = (x + y)^z$.
- (h) $f(x, y) = \sin(xy)$.
- (i) $f(x, y) = \{\sin(xy)\}^{\cos z}$.

2-18. Find the partial derivatives of the following functions (where $g: \mathbf{R} \rightarrow \mathbf{R}$ is continuous):

- (a) $f(x, y) = \int_a^{x+y} g$.
- (b) $f(x, y) = \int_y^x g$.
- (c) $f(x, y) = \int_a^{xy} g$.
- (d) $f(x, y) = \int_a^{\left(\int_b^y g\right)} g$.

2-19. If $f(x, y) = x^{x^{xy}} + (\log x)(\arctan(\arctan(\arctan(\sin(\cos xy) - \log(x + y))))))$ find $D_2f(1, y)$. *Hint:* There is an easy way to do this.

2-20. Find the partial derivatives of f in terms of the derivatives of g and h if

- (a) $f(x, y) = g(x)h(y)$.
- (b) $f(x, y) = g(x)^{h(y)}$.
- (c) $f(x, y) = g(x)$.
- (d) $f(x, y) = g(y)$.
- (e) $f(x, y) = g(x + y)$.

2-21.* Let $g_1, g_2: \mathbf{R}^2 \rightarrow \mathbf{R}$ be continuous. Define $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ by

$$f(x, y) = \int_0^x g_1(t, 0) dt + \int_0^y g_2(x, t) dt.$$

- (a) Show that $D_2f(x, y) = g_2(x, y)$.
- (b) How should f be defined so that $D_1f(x, y) = g_1(x, y)$?
- (c) Find a function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ such that $D_1f(x, y) = x$ and $D_2f(x, y) = y$. Find one such that $D_1f(x, y) = y$ and $D_2f(x, y) = x$.

2-22.* If $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ and $D_2f = 0$, show that f is independent of the second variable. If $D_1f = D_2f = 0$, show that f is constant.

2-23.* Let $A = \{(x, y) \in \mathbf{R}^2: x < 0, \text{ or } x \geq 0 \text{ and } y \neq 0\}$.

(a) If $f: A \rightarrow \mathbf{R}$ and $D_1f = D_2f = 0$, show that f is constant.

Hint: Note that any two points in A can be connected by a sequence of lines each parallel to one of the axes.

(b) Find a function $f: A \rightarrow \mathbf{R}$ such that $D_2f = 0$ but f is not independent of the second variable.

2-24. Define $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ by

$$f(x, y) = \begin{cases} xy \frac{x^2 - y^2}{x^2 + y^2} & (x, y) \neq 0, \\ 0 & (x, y) = 0. \end{cases}$$

(a) Show that $D_2f(x, 0) = x$ for all x and $D_1f(0, y) = -y$ for all y .

(b) Show that $D_{1,2}f(0, 0) \neq D_{2,1}f(0, 0)$.

2-25.* Define $f: \mathbf{R} \rightarrow \mathbf{R}$ by

$$f(x) = \begin{cases} e^{-x^{-2}} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Show that f is a C^∞ function, and $f^{(i)}(0) = 0$ for all i . *Hint:*

The limit $f'(0) = \lim_{h \rightarrow 0} \frac{e^{-h^{-2}}}{h} = \lim_{h \rightarrow 0} \frac{1/h}{e^{h^{-2}}}$ can be evaluated by

L'Hospital's rule. It is easy enough to find $f'(x)$ for $x \neq 0$, and $f''(0) = \lim_{h \rightarrow 0} f'(h)/h$ can then be found by L'Hospital's rule.

2-26.* Let $f(x) = \begin{cases} e^{-(x-1)^{-2}} \cdot e^{-(x+1)^{-2}} & x \in (-1, 1), \\ 0 & x \notin (-1, 1). \end{cases}$

(a) Show that $f: \mathbf{R} \rightarrow \mathbf{R}$ is a C^∞ function which is positive on $(-1, 1)$ and 0 elsewhere.

(b) Show that there is a C^∞ function $g: \mathbf{R} \rightarrow [0, 1]$ such that $g(x) = 0$ for $x \leq 0$ and $g(x) = 1$ for $x \geq \varepsilon$. *Hint:* If f is a C^∞ function which is positive on $(0, \varepsilon)$ and 0 elsewhere, let $g(x) = \int_0^x f / \int_0^\varepsilon f$.

(c) If $a \in \mathbf{R}^n$, define $g: \mathbf{R}^n \rightarrow \mathbf{R}$ by

$$g(x) = f([x^1 - a^1]/\varepsilon) \cdot \dots \cdot f([x^n - a^n]/\varepsilon).$$

Show that g is a C^∞ function which is positive on

$$(a^1 - \varepsilon, a^1 + \varepsilon) \times \dots \times (a^n - \varepsilon, a^n + \varepsilon)$$

and zero elsewhere.

(d) If $A \subset \mathbf{R}^n$ is open and $C \subset A$ is compact, show that there is a non-negative C^∞ function $f: A \rightarrow \mathbf{R}$ such that $f(x) > 0$ for $x \in C$ and $f = 0$ outside of some closed set contained in A .

(e) Show that we can choose such an f so that $f: A \rightarrow [0, 1]$ and $f(x) = 1$ for $x \in C$. *Hint:* If the function f of (d) satisfies $f(x) \geq \varepsilon$ for $x \in C$, consider $g \circ f$, where g is the function of (b).

2-27. Define $g, h: \{x \in \mathbf{R}^2: |x| \leq 1\} \rightarrow \mathbf{R}^3$ by

$$\begin{aligned} g(x, y) &= (x, y, \sqrt{1 - x^2 - y^2}), \\ h(x, y) &= (x, y, -\sqrt{1 - x^2 - y^2}). \end{aligned}$$

Show that the maximum of f on $\{x \in \mathbf{R}^2: |x| = 1\}$ is either the maximum of $f \circ g$ or the maximum of $f \circ h$ on $\{x \in \mathbf{R}^2: |x| \leq 1\}$.

DERIVATIVES

The reader who has compared Problems 2-10 and 2-17 has probably already guessed the following.

2-7 Theorem. *If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable at a , then $D_j f^i(a)$ exists for $1 \leq i \leq m$, $1 \leq j \leq n$ and $f'(a)$ is the $m \times n$ matrix $(D_j f^i(a))$.*

Proof. Suppose first that $m = 1$, so that $f: \mathbf{R}^n \rightarrow \mathbf{R}$. Define $h: \mathbf{R} \rightarrow \mathbf{R}^n$ by $h(x) = (a^1, \dots, x, \dots, a^n)$, with x in the j th place. Then $D_j f(a) = (f \circ h)'(a^j)$. Hence, by Theorem 2-2,

$$\begin{aligned} (f \circ h)'(a^j) &= f'(a) \cdot h'(a^j) \\ &= f'(a) \cdot \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j\text{th place.} \end{aligned}$$

Since $(f \circ h)'(a^j)$ has the single entry $D_j f(a)$, this shows that $D_j f(a)$ exists and is the j th entry of the $1 \times n$ matrix $f'(a)$.

The theorem now follows for arbitrary m since, by Theorem 2-3, each f^i is differentiable and the i th row of $f'(a)$ is $(f^i)'(a)$. ■

There are several examples in the problems to show that the converse of Theorem 2-7 is false. It is true, however, if one hypothesis is added.

2-8 Theorem. If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, then $Df(a)$ exists if all $D_j f^i(x)$ exist in an open set containing a and if each function $D_j f^i$ is continuous at a .

(Such a function f is called **continuously differentiable** at a .)

Proof. As in the proof of Theorem 2-7, it suffices to consider the case $m = 1$, so that $f: \mathbf{R}^n \rightarrow \mathbf{R}$. Then

$$\begin{aligned} f(a + h) - f(a) &= f(a^1 + h^1, a^2, \dots, a^n) - f(a^1, \dots, a^n) \\ &\quad + f(a^1 + h^1, a^2 + h^2, a^3, \dots, a^n) \\ &\quad \quad - f(a^1 + h^1, a^2, \dots, a^n) \\ &\quad + \dots \\ &\quad + f(a^1 + h^1, \dots, a^{n-1} + h^{n-1}, a^n) \\ &\quad \quad - f(a^1 + h^1, \dots, a^{n-1}, a^n). \end{aligned}$$

Recall that $D_1 f$ is the derivative of the function g defined by $g(x) = f(x, a^2, \dots, a^n)$. Applying the mean-value theorem to g we obtain

$$\begin{aligned} f(a^1 + h^1, a^2, \dots, a^n) - f(a^1, \dots, a^n) \\ = h^1 \cdot D_1 f(b_1, a^2, \dots, a^n) \end{aligned}$$

for some b_1 between a^1 and $a^1 + h^1$. Similarly the i th term in the sum equals

$$h^i \cdot D_i f(a^1 + h^1, \dots, a^{i-1} + h^{i-1}, b_i, \dots, a^n) = h^i D_i f(c_i),$$

for some c_i . Then

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\left| f(a + h) - f(a) - \sum_{i=1}^n D_i f(a) \cdot h^i \right|}{|h|} \\ = \lim_{h \rightarrow 0} \frac{\left| \sum_{i=1}^n [D_i f(c_i) - D_i f(a)] \cdot h^i \right|}{|h|} \\ \leq \lim_{h \rightarrow 0} \sum_{i=1}^n |D_i f(c_i) - D_i f(a)| \cdot \frac{|h^i|}{|h|} \\ \leq \lim_{h \rightarrow 0} \sum_{i=1}^n |D_i f(c_i) - D_i f(a)| \\ = 0, \end{aligned}$$

since $D_i f$ is continuous at a . ■

Although the chain rule was used in the proof of Theorem 2-7, it could easily have been eliminated. With Theorem 2-8 to provide differentiable functions, and Theorem 2-7 to provide their derivatives, the chain rule may therefore seem almost superfluous. However, it has an extremely important corollary concerning partial derivatives.

2-9 Theorem. Let $g_1, \dots, g_m: \mathbf{R}^n \rightarrow \mathbf{R}$ be continuously differentiable at a , and let $f: \mathbf{R}^m \rightarrow \mathbf{R}$ be differentiable at $(g_1(a), \dots, g_m(a))$. Define the function $F: \mathbf{R}^n \rightarrow \mathbf{R}$ by $F(x) = f(g_1(x), \dots, g_m(x))$. Then

$$D_i F(a) = \sum_{j=1}^m D_j f(g_1(a), \dots, g_m(a)) \cdot D_i g_j(a).$$

Proof. The function F is just the composition $f \circ g$, where $g = (g_1, \dots, g_m)$. Since g_i is continuously differentiable at a , it follows from Theorem 2-8 that g is differentiable at a . Hence by Theorem 2-2,

$$F'(a) = f'(g(a)) \cdot g'(a) = (D_1 f(g(a)), \dots, D_m f(g(a))) \cdot \begin{pmatrix} D_1 g_1(a), & \dots, & D_n g_1(a) \\ \vdots & & \vdots \\ D_1 g_m(a), & \dots, & D_n g_m(a) \end{pmatrix}$$

But $D_i F(a)$ is the i th entry of the left side of this equation, while $\sum_{j=1}^m D_j f(g_1(a), \dots, g_m(a)) \cdot D_i g_j(a)$ is the i th entry of the right side. ■

Theorem 2-9 is often called the *chain rule*, but is weaker than Theorem 2-2 since g could be differentiable without g_i being continuously differentiable (see Problem 2-32). Most computations requiring Theorem 2-9 are fairly straightforward. A slight subtlety is required for the function $F: \mathbf{R}^2 \rightarrow \mathbf{R}$ defined by

$$F(x, y) = f(g(x, y), h(x), k(y))$$

where $h, k: \mathbf{R} \rightarrow \mathbf{R}$. In order to apply Theorem 2-9 define $\bar{h}, \bar{k}: \mathbf{R}^2 \rightarrow \mathbf{R}$ by

$$\bar{h}(x, y) = h(x) \quad \bar{k}(x, y) = k(y).$$

Then

$$\begin{aligned} D_1 \bar{h}(x, y) &= h'(x) & D_2 \bar{h}(x, y) &= 0, \\ D_1 \bar{k}(x, y) &= 0 & D_2 \bar{k}(x, y) &= k'(y), \end{aligned}$$

and we can write

$$F(x, y) = f(g(x, y), \bar{h}(x, y), \bar{k}(x, y)).$$

Letting $a = (g(x, y), h(x), k(y))$, we obtain

$$\begin{aligned} D_1 F(x, y) &= D_1 f(a) \cdot D_1 g(x, y) + D_2 f(a) \cdot h'(x), \\ D_2 F(x, y) &= D_1 f(a) \cdot D_2 g(x, y) + D_3 f(a) \cdot k'(y). \end{aligned}$$

It should, of course, be unnecessary for you to actually write down the functions \bar{h} and \bar{k} .

Problems. 2-28. Find expressions for the partial derivatives of the following functions:

- (a) $F(x, y) = f(g(x)k(y), g(x) + h(y))$.
- (b) $F(x, y, z) = f(g(x + y), h(y + z))$.
- (c) $F(x, y, z) = f(x^y, y^z, z^x)$.
- (d) $F(x, y) = f(x, g(x), h(x, y))$.

2-29. Let $f: \mathbf{R}^n \rightarrow \mathbf{R}$. For $x \in \mathbf{R}^n$, the limit

$$\lim_{t \rightarrow 0} \frac{f(a + tx) - f(a)}{t},$$

if it exists, is denoted $D_x f(a)$, and called the **directional derivative** of f at a , in the direction x .

- (a) Show that $D_{e_i} f(a) = D_i f(a)$.
 - (b) Show that $D_{tx} f(a) = t D_x f(a)$.
 - (c) If f is differentiable at a , show that $D_x f(a) = Df(a)(x)$ and therefore $D_{x+y} f(a) = D_x f(a) + D_y f(a)$.
- 2-30.** Let f be defined as in Problem 2-4. Show that $D_x f(0, 0)$ exists for all x , but if $g \neq 0$, then $D_{x+y} f(0, 0) = D_x f(0, 0) + D_y f(0, 0)$ is not true for all x and y .
- 2-31.** Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined as in Problem 1-26. Show that $D_x f(0, 0)$ exists for all x , although f is not even continuous at $(0, 0)$.
- 2-32.** (a) Let $f: \mathbf{R} \rightarrow \mathbf{R}$ be defined by

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Show that f is differentiable at 0 but f' is not continuous at 0.

(b) Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} (x^2 + y^2) \sin \frac{1}{\sqrt{x^2 + y^2}} & (x, y) \neq 0, \\ 0 & (x, y) = 0. \end{cases}$$

Show that f is differentiable at $(0, 0)$ but $D_i f$ is not continuous at $(0, 0)$.

2-33. Show that the continuity of $D_1 f^j$ at a may be eliminated from the hypothesis of Theorem 2-8.

2-34. A function $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is **homogeneous** of degree m if $f(tx) = t^m f(x)$ for all x . If f is also differentiable, show that

$$\sum_{i=1}^n x^i D_i f(x) = m f(x).$$

Hint: If $g(t) = f(tx)$, find $g'(1)$.

2-35. If $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable and $f(0) = 0$, prove that there exist $g_i: \mathbf{R}^n \rightarrow \mathbf{R}$ such that

$$f(x) = \sum_{i=1}^n x^i g_i(x).$$

Hint: If $h_x(t) = f(tx)$, then $f(x) = \int_0^1 h_x'(t) dt$.

INVERSE FUNCTIONS

Suppose that $f: \mathbf{R} \rightarrow \mathbf{R}$ is continuously differentiable in an open set containing a and $f'(a) \neq 0$. If $f'(a) > 0$, there is an open interval V containing a such that $f'(x) > 0$ for $x \in V$, and a similar statement holds if $f'(a) < 0$. Thus f is increasing (or decreasing) on V , and is therefore 1-1 with an inverse function f^{-1} defined on some open interval W containing $f(a)$. Moreover it is not hard to show that f^{-1} is differentiable, and for $y \in W$ that

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}.$$

An analogous discussion in higher dimensions is much more involved, but the result (Theorem 2-11) is very important. We begin with a simple lemma.

2-10 Lemma. Let $A \subset \mathbf{R}^n$ be a rectangle and let $f: A \rightarrow \mathbf{R}^n$ be continuously differentiable. If there is a number M such that $|D_j f^i(x)| \leq M$ for all x in the interior of A , then

$$|f(x) - f(y)| \leq n^2 M |x - y|$$

for all $x, y \in A$.

Proof. We have

$$\begin{aligned} f^i(y) - f^i(x) &= \sum_{j=1}^n [f^i(y^1, \dots, y^j, x^{j+1}, \dots, x^n) \\ &\quad - f^i(y^1, \dots, y^{j-1}, x^j, \dots, x^n)]. \end{aligned}$$

Applying the mean-value theorem we obtain

$$\begin{aligned} f^i(y^1, \dots, y^j, x^{j+1}, \dots, x^n) - f^i(y^1, \dots, y^{j-1}, x^j, \dots, x^n) \\ = (y^j - x^j) \cdot D_j f^i(z_{ij}) \end{aligned}$$

for some z_{ij} . The expression on the right has absolute value less than or equal to $M \cdot |y^j - x^j|$. Thus

$$|f^i(y) - f^i(x)| \leq \sum_{j=1}^n |y^j - x^j| \cdot M \leq nM |y - x|$$

since each $|y^j - x^j| \leq |y - x|$. Finally

$$|f(y) - f(x)| \leq \sum_{i=1}^n |f^i(y) - f^i(x)| \leq n^2 M \cdot |y - x|. \quad \blacksquare$$

2-11 Theorem (Inverse Function Theorem). Suppose that $f: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is continuously differentiable in an open set containing a , and $\det f'(a) \neq 0$. Then there is an open set V containing a and an open set W containing $f(a)$ such that $f: V \rightarrow W$ has a continuous inverse $f^{-1}: W \rightarrow V$ which is differentiable and for all $y \in W$ satisfies

$$(f^{-1})'(y) = [f'(f^{-1}(y))]^{-1}.$$

Proof. Let λ be the linear transformation $Df(a)$. Then λ is non-singular, since $\det f'(a) \neq 0$. Now $D(\lambda^{-1} \circ f)(a) = D(\lambda^{-1})(f(a)) \circ Df(a) = \lambda^{-1} \circ Df(a)$ is the identity linear

transformation. If the theorem is true for $\lambda^{-1} \circ f$, it is clearly true for f . Therefore we may assume at the outset that λ is the identity. Thus whenever $f(a + h) = f(a)$, we have

$$\frac{|f(a + h) - f(a) - \lambda(h)|}{|h|} = \frac{|h|}{|h|} = 1.$$

But

$$\lim_{h \rightarrow 0} \frac{|f(a + h) - f(a) - \lambda(h)|}{|h|} = 0.$$

This means that we cannot have $f(x) = f(a)$ for x arbitrarily close to, but unequal to, a . Therefore there is a closed rectangle U containing a in its interior such that

1. $f(x) \neq f(a)$ if $x \in U$ and $x \neq a$.

Since f is continuously differentiable in an open set containing a , we can also assume that

2. $\det f'(x) \neq 0$ for $x \in U$.
3. $|D_j f^i(x) - D_j f^i(a)| < 1/2n^2$ for all i, j , and $x \in U$.

Note that (3) and Lemma 2-10 applied to $g(x) = f(x) - x$ imply for $x_1, x_2 \in U$ that

$$|f(x_1) - x_1 - (f(x_2) - x_2)| \leq \frac{1}{2}|x_1 - x_2|.$$

Since

$$\begin{aligned} |x_1 - x_2| - |f(x_1) - f(x_2)| &\leq |f(x_1) - x_1 - (f(x_2) - x_2)| \\ &\leq \frac{1}{2}|x_1 - x_2|, \end{aligned}$$

we obtain

4. $|x_1 - x_2| \leq 2|f(x_1) - f(x_2)|$ for $x_1, x_2 \in U$.

Now $f(\text{boundary } U)$ is a compact set which, by (1), does not contain $f(a)$ (Figure 2-3). Therefore there is a number $d > 0$ such that $|f(a) - f(x)| \geq d$ for $x \in \text{boundary } U$. Let $W = \{y: |y - f(a)| < d/2\}$. If $y \in W$ and $x \in \text{boundary } U$, then

5. $|y - f(a)| < |y - f(x)|$.

We will show that for any $y \in W$ there is a unique x in interior U such that $f(x) = y$. To prove this consider the

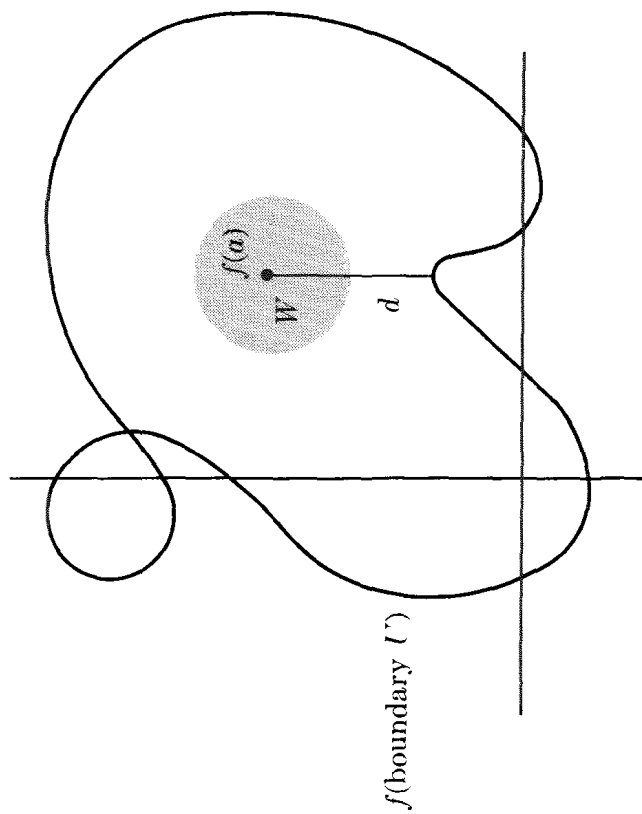


FIGURE 2-3

function $g: U \rightarrow \mathbf{R}$ defined by

$$g(x) = |y - f(x)|^2 = \sum_{i=1}^n (y^i - f^i(x))^2.$$

This function is continuous and therefore has a minimum on U . If $x \in \text{boundary } U$, then, by (5), we have $g(a) < g(x)$. Therefore the minimum of g does *not* occur on the boundary of U . By Theorem 2-6 there is a point $x \in \text{interior } U$ such that $D_j g(x) = 0$ for all j , that is

$$\sum_{i=1}^n 2(y^i - f^i(x)) \cdot D_j f^i(x) = 0 \quad \text{for all } j.$$

By (2) the matrix $(D_j f^i(x))$ has non-zero determinant. Therefore we must have $y^i - f^i(x) = 0$ for all i , that is $y = f(x)$. This proves the existence of x . Uniqueness follows immediately from (4).

If $V = (\text{interior } U) \cap f^{-1}(W)$, we have shown that the function $f: V \rightarrow W$ has an inverse $f^{-1}: W \rightarrow V$. We can rewrite (4) as

$$6. |f^{-1}(y_1) - f^{-1}(y_2)| \leq 2|y_1 - y_2| \quad \text{for } y_1, y_2 \in W.$$

This shows that f^{-1} is continuous.

Only the proof that f^{-1} is differentiable remains. Let $\mu = Df(x)$. We will show that f^{-1} is differentiable at $y = f(x)$ with derivative μ^{-1} . As in the proof of Theorem 2-2, for $x_1 \in V$, we have

$$f(x_1) = f(x) + \mu(x_1 - x) + \varphi(x_1 - x),$$

where

$$\lim_{x_1 \rightarrow x} \frac{|\varphi(x_1 - x)|}{|x_1 - x|} = 0.$$

Therefore

$$\mu^{-1}(f(x_1) - f(x)) = x_1 - x + \mu^{-1}(\varphi(x_1 - x)).$$

Since every $y_1 \in W$ is of the form $f(x_1)$ for some $x_1 \in V$, this can be written

$$f^{-1}(y_1) = f^{-1}(y) + \mu^{-1}(y_1 - y) - \mu^{-1}(\varphi(f^{-1}(y_1) - f^{-1}(y))),$$

and it therefore suffices to show that

$$\lim_{y_1 \rightarrow y} \frac{|\mu^{-1}(\varphi(f^{-1}(y_1) - f^{-1}(y)))|}{|y_1 - y|} = 0.$$

Therefore (Problem 1-10) it suffices to show that

$$\lim_{y_1 \rightarrow y} \frac{|\varphi(f^{-1}(y_1) - f^{-1}(y))|}{|y_1 - y|} = 0.$$

Now

$$\begin{aligned} \frac{|\varphi(f^{-1}(y_1) - f^{-1}(y))|}{|y_1 - y|} &= \frac{|\varphi(f^{-1}(y_1) - f^{-1}(y))|}{|f^{-1}(y_1) - f^{-1}(y)|} \cdot \frac{|f^{-1}(y_1) - f^{-1}(y)|}{|y_1 - y|}. \end{aligned}$$

Since f^{-1} is continuous, $f^{-1}(y_1) \rightarrow f^{-1}(y)$ as $y_1 \rightarrow y$. Therefore the first factor approaches 0. Since, by (6), the second factor is less than 2, the product also approaches 0. ■

It should be noted that an inverse function f^{-1} may exist even if $\det f'(a) = 0$. For example, if $f: \mathbf{R} \rightarrow \mathbf{R}$ is defined by $f(x) = x^3$, then $f'(0) = 0$ but f has the inverse function $f^{-1}(x) = \sqrt[3]{x}$. One thing is certain however: if $\det f'(a) = 0$, then f^{-1} cannot be differentiable at $f(a)$. To prove this note that $f \circ f^{-1}(x) = x$. If f^{-1} were differentiable at $f(a)$, the chain rule would give $f'(a) \cdot (f^{-1})'(f(a)) = I$, and consequently $\det f'(a) \cdot \det(f^{-1})'(f(a)) = 1$, contradicting $\det f'(a) = 0$.

Problems. 2-36.* Let $A \subset \mathbf{R}^n$ be an open set and $f: A \rightarrow \mathbf{R}^n$ a continuously differentiable 1-1 function such that $\det f'(x) \neq 0$ for all x . Show that $f(A)$ is an open set and $f^{-1}: f(A) \rightarrow A$ is differentiable. Show also that $f(B)$ is open for any open set $B \subset A$.

2-37. (a) Let $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ be a continuously differentiable function. Show that f is *not* 1-1. *Hint:* If, for example, $D_1 f(x, y) \neq 0$ for all (x, y) in some open set A , consider $g: A \rightarrow \mathbf{R}^2$ defined by $g(x, y) = (f(x, y), y)$.

(b) Generalize this result to the case of a continuously differentiable function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ with $m < n$.

2-38. (a) If $f: \mathbf{R} \rightarrow \mathbf{R}$ satisfies $f'(a) \neq 0$ for all $a \in \mathbf{R}$, show that f is 1-1 (on all of \mathbf{R}).

(b) Define $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$ by $f(x,y) = (e^x \cos y, e^x \sin y)$. Show that $\det f'(x,y) \neq 0$ for all (x,y) but f is not 1-1.

2-39. Use the function $f: \mathbf{R} \rightarrow \mathbf{R}$ defined by

$$f(x) = \begin{cases} \frac{x}{2} + x^2 \sin \frac{1}{x} & x \neq 0, \\ 0 & x = 0, \end{cases}$$

to show that continuity of the derivative cannot be eliminated from the hypothesis of Theorem 2-11.

IMPLICIT FUNCTIONS

Consider the function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ defined by $f(x,y) = x^2 + y^2 - 1$. If we choose (a,b) with $f(a,b) = 0$ and $a \neq 1, -1$, there are (Figure 2-4) open intervals A containing a and B containing b with the following property: if $x \in A$, there is a unique $y \in B$ with $f(x,y) = 0$. We can therefore define

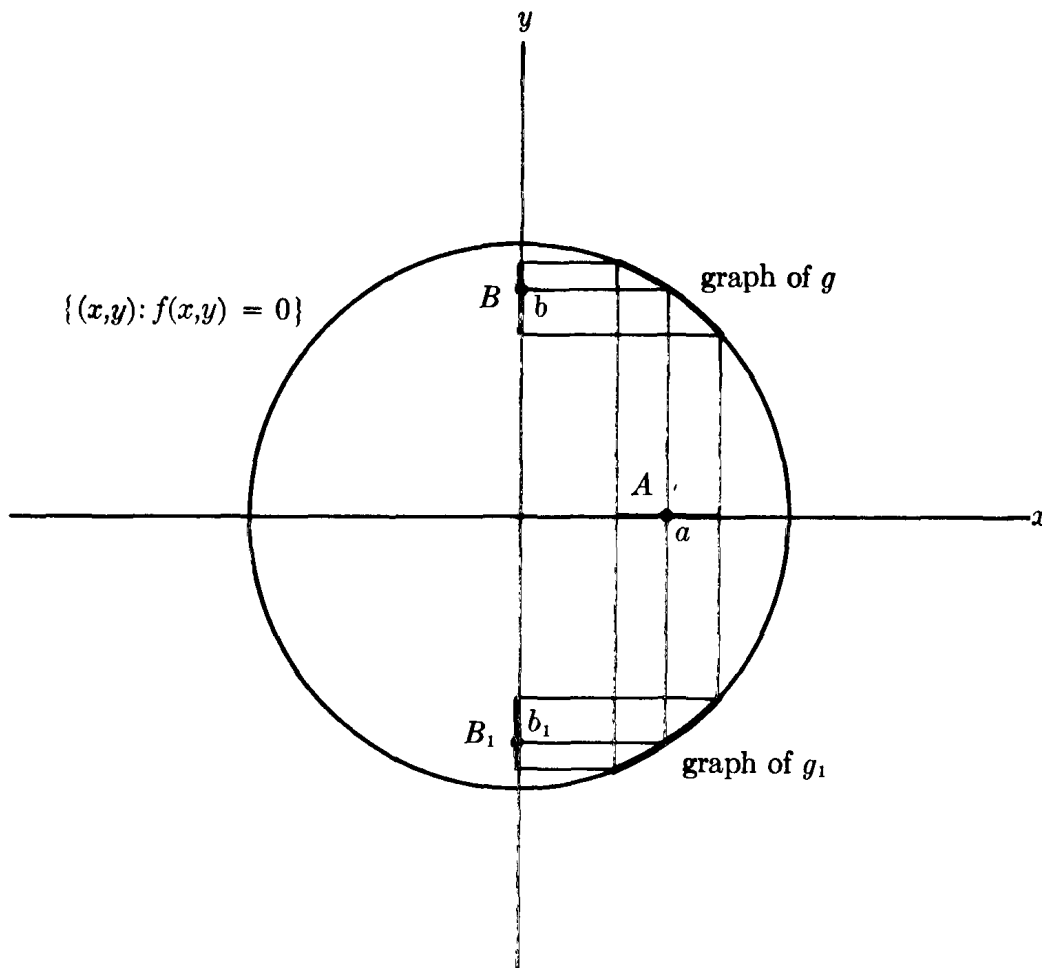


FIGURE 2-4

a function $g: A \rightarrow \mathbf{R}$ by the condition $g(x) \in B$ and $f(x, g(x)) = 0$ (if $b > 0$, as indicated in Figure 2-4, then $g(x) = \sqrt{1 - x^2}$). For the function f we are considering there is another number b_1 such that $f(a, b_1) = 0$. There will also be an interval B_1 containing b_1 such that, when $x \in A$, we have $f(x, g_1(x)) = 0$ for a unique $g_1(x) \in B_1$ (here $g_1(x) = -\sqrt{1 - x^2}$). Both g and g_1 are differentiable. These functions are said to be defined **implicitly** by the equation $f(x, y) = 0$.

If we choose $a = 1$ or -1 it is impossible to find any such function g defined in an open interval containing a . We would like a simple criterion for deciding when, in general, such a function can be found. More generally we may ask the following: If $f: \mathbf{R}^n \times \mathbf{R} \rightarrow \mathbf{R}$ and $f(a^1, \dots, a^n, b) = 0$, when can we find, for each (x^1, \dots, x^n) near (a^1, \dots, a^n) , a unique y near b such that $f(x^1, \dots, x^n, y) = 0$? Even more generally, we can ask about the possibility of solving m equations, depending upon parameters x^1, \dots, x^n , in m unknowns: If

$$f_i: \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R} \quad i = 1, \dots, m$$

and

$$f_i(a^1, \dots, a^n, b^1, \dots, b^m) = 0 \quad i = 1, \dots, m,$$

when can we find, for each (x^1, \dots, x^n) near (a^1, \dots, a^n) a unique (y^1, \dots, y^m) near (b^1, \dots, b^m) which satisfies $f_i(x^1, \dots, x^n, y^1, \dots, y^m) = 0$? The answer is provided by

2-12 Theorem (Implicit Function Theorem). Suppose $f: \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}^m$ is continuously differentiable in an open set containing (a, b) and $f(a, b) = 0$. Let M be the $m \times m$ matrix

$$(D_{n+j}f^i(a, b)) \quad 1 \leq i, j \leq m.$$

If $\det M \neq 0$, there is an open set $A \subset \mathbf{R}^n$ containing a and an open set $B \subset \mathbf{R}^m$ containing b , with the following property: for each $x \in A$ there is a unique $g(x) \in B$ such that $f(x, g(x)) = 0$. The function g is differentiable.

Proof. Define $F: \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}^n \times \mathbf{R}^m$ by $F(x, y) = (x, f(x, y))$. Then $\det F'(a, b) = \det M \neq 0$. By Theorem 2-11 there is an open set $W \subset \mathbf{R}^n \times \mathbf{R}^m$ containing $F(a, b) = (a, 0)$ and an open set in $\mathbf{R}^n \times \mathbf{R}^m$ containing (a, b) , which we may take to be of the form $A \times B$, such that $F: A \times B \rightarrow W$ has a differentiable inverse $h: W \rightarrow A \times B$. Clearly h is of the form $h(x, y) = (x, k(x, y))$ for some differentiable function k (since F is of this form). Let $\pi: \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}^m$ be defined by $\pi(x, y) = y$; then $\pi \circ F = f$. Therefore

$$\begin{aligned} f(x, k(x, y)) &= f \circ h(x, y) = (\pi \circ F) \circ h(x, y) \\ &= \pi \circ (F \circ h)(x, y) = \pi(x, y) = y. \end{aligned}$$

Thus $f(x, k(x, 0)) = 0$; in other words we can define $g(x) = k(x, 0)$. ■

Since the function g is known to be differentiable, it is easy to find its derivative. In fact, since $f^i(x, g(x)) = 0$, taking D_j of both sides gives

$$0 = D_j f^i(x, g(x)) + \sum_{\alpha=1}^m D_{n+\alpha} f^i(x, g(x)) \cdot D_j g^\alpha(x)$$

$$i, j = 1, \dots, m.$$

Since $\det M \neq 0$, these equations can be solved for $D_j g^\alpha(x)$. The answer will depend on the various $D_j f^i(x, g(x))$, and therefore on $g(x)$. This is unavoidable, since the function g is not unique. Reconsidering the function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ defined by $f(x, y) = x^2 + y^2 - 1$, we note that two possible functions satisfying $f(x, g(x)) = 0$ are $g(x) = \sqrt{1 - x^2}$ and $g(x) = -\sqrt{1 - x^2}$. Differentiating $f(x, g(x)) = 0$ gives

$$D_1 f(x, g(x)) + D_2 f(x, g(x)) \cdot g'(x) = 0,$$

or

$$\begin{aligned} 2x + 2g(x) \cdot g'(x) &= 0, \\ g'(x) &= -x/g(x), \end{aligned}$$

which is indeed the case for either $g(x) = \sqrt{1 - x^2}$ or $g(x) = -\sqrt{1 - x^2}$.

A generalization of the argument for Theorem 2-12 can be given, which will be important in Chapter 5.

2-13 Theorem. Let $f: \mathbf{R}^n \rightarrow \mathbf{R}^p$ be continuously differentiable in an open set containing a , where $p \leq n$. If $f(a) = 0$ and the $p \times n$ matrix $(D_j f^i(a))$ has rank p , then there is an open set $A \subset \mathbf{R}^n$ containing a and a differentiable function $h: A \rightarrow \mathbf{R}^n$ with differentiable inverse such that

$$f \circ h(x^1, \dots, x^n) = (x^{n-p+1}, \dots, x^n).$$

Proof. We can consider f as a function $f: \mathbf{R}^{n-p} \times \mathbf{R}^p \rightarrow \mathbf{R}^p$. If $\det M \neq 0$, then M is the $p \times p$ matrix $(D_{n-p+j} f^i(a))$, $1 \leq i, j \leq p$, then we are precisely in the situation considered in the proof of Theorem 2-12, and as we showed in that proof, there is h such that $f \circ h(x^1, \dots, x^n) = (x^{n-p+1}, \dots, x^n)$.

In general, since $(D_j f^i(a))$ has rank p , there will be $j_1 < \dots < j_p$ such that the matrix $(D_j f^i(a))$ $1 \leq i \leq p$, $j = j_1, \dots, j_p$ has non-zero determinant. If $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$ permutes the x^j so that $g(x^1, \dots, x^n) = (\dots, x^{j_1}, \dots, x^{j_p})$, then $f \circ g$ is a function of the type already considered, so $((f \circ g) \circ k)(x^1, \dots, x^n) = (x^{n-p+1}, \dots, x^n)$ for some k . Let $h = g \circ k$. ■

Problems. 2-40. Use the implicit function theorem to re-do Problem 2-15(c).

2-41. Let $f: \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$ be differentiable. For each $x \in \mathbf{R}$ define $g_x: \mathbf{R} \rightarrow \mathbf{R}$ by $g_x(y) = f(x, y)$. Suppose that for each x there is a unique y with $g'_x(y) = 0$; let $c(x)$ be this y .

(a) If $D_{2,2}f(x, y) \neq 0$ for all (x, y) , show that c is differentiable and

$$c'(x) = - \frac{D_{2,1}f(x, c(x))}{D_{2,2}f(x, c(x))}.$$

Hint: $g'_x(y) = 0$ can be written $D_2f(x, y) = 0$.

(b) Show that if $c'(x) = 0$, then for some y we have

$$\begin{aligned} D_{2,1}f(x, y) &= 0, \\ D_{2,2}f(x, y) &= 0. \end{aligned}$$

(c) Let $f(x, y) = x(y \log y - y) - y \log x$. Find

$$\max_{\frac{1}{2} \leq x \leq 2} \left(\min_{\frac{1}{2} \leq y \leq 1} f(x, y) \right).$$

NOTATION

This section is a brief and not entirely unprejudiced discussion of classical notation connected with partial derivatives.

The partial derivative $D_1f(x,y,z)$ is denoted, among devotees of classical notation, by

$$\frac{\partial f(x,y,z)}{\partial x} \quad \text{or} \quad \frac{\partial f}{\partial x} \quad \text{or} \quad \frac{\partial f}{\partial x}(x,y,z) \quad \text{or} \quad \frac{\partial}{\partial x} f(x,y,z)$$

or any other convenient similar symbol. This notation forces one to write

$$\frac{\partial f}{\partial u}(u,v,w)$$

for $D_1f(u,v,w)$, although the symbol

$$\left. \frac{\partial f(x,y,z)}{\partial x} \right|_{(x,y,z)=(u,v,w)} \quad \text{or} \quad \frac{\partial f(x,y,z)}{\partial x}(u,v,w)$$

or something similar may be used (and must be used for an expression like $D_1f(7,3,2)$). Similar notation is used for D_2f and D_3f . Higher-order derivatives are denoted by symbols like

$$D_2D_1f(x,y,z) = \frac{\partial^2 f(x,y,z)}{\partial y \partial x}.$$

When $f: \mathbf{R} \rightarrow \mathbf{R}$, the symbol ∂ automatically reverts to d ; thus

$$\frac{d \sin x}{dx}, \quad \text{not} \quad \frac{\partial \sin x}{\partial x}.$$

The mere statement of Theorem 2-2 in classical notation requires the introduction of irrelevant letters. The usual evaluation for $D_1(f \circ (g,h))$ runs as follows:

If $f(u,v)$ is a function and $u = g(x,y)$ and $v = h(x,y)$, then

$$\frac{\partial f(g(x,y), h(x,y))}{\partial x} = \frac{\partial f(u,v)}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial f(u,v)}{\partial v} \frac{\partial v}{\partial x}.$$

[The symbol $\partial u / \partial x$ means $\partial / \partial x g(x,y)$ and $\partial / \partial u f(u,v)$ means

$D_1f(u,v) = D_1f(g(x,y), h(x,y)).]$ This equation is often written simply

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial x}.$$

Note that f means something different on the two sides of the equation!

The notation df/dx , always a little too tempting, has inspired many (usually meaningless) definitions of dx and df separately, the sole purpose of which is to make the equation

$$df = \frac{df}{dx} \cdot dx$$

work out. If $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ then df is *defined*, classically, as

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy$$

(whatever dx and dy mean).

Chapter 4 contains rigorous definitions which enable us to prove the above equations as theorems. It is a touchy question whether or not these modern definitions represent a real improvement over classical formalism; this the reader must decide for himself.

3

Integration

BASIC DEFINITIONS

The definition of the integral of a function $f: A \rightarrow \mathbf{R}$, where $A \subset \mathbf{R}^n$ is a closed rectangle, is so similar to that of the ordinary integral that a rapid treatment will be given.

Recall that a partition P of a closed interval $[a, b]$ is a sequence t_0, \dots, t_k , where $a = t_0 \leq t_1 \leq \dots \leq t_k = b$. The partition P divides the interval $[a, b]$ into k subintervals $[t_{i-1}, t_i]$. A **partition** of a rectangle $[a_1, b_1] \times \dots \times [a_n, b_n]$ is a collection $P = (P_1, \dots, P_n)$, where each P_i is a partition of the interval $[a_i, b_i]$. Suppose, for example, that $P_1 = t_0, \dots, t_k$ is a partition of $[a_1, b_1]$ and $P_2 = s_0, \dots, s_l$ is a partition of $[a_2, b_2]$. Then the partition $P = (P_1, P_2)$ of $[a_1, b_1] \times [a_2, b_2]$ divides the closed rectangle $[a_1, b_1] \times [a_2, b_2]$ into $k \cdot l$ subrectangles, a typical one being $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$. In general, if P_i divides $[a_i, b_i]$ into N_i subintervals, then $P = (P_1, \dots, P_n)$ divides $[a_1, b_1] \times \dots \times [a_n, b_n]$ into $N = N_1 \cdot \dots \cdot N_n$ subrectangles. These subrectangles will be called **subrectangles of the partition P** .

Suppose now that A is a rectangle, $f: A \rightarrow \mathbf{R}$ is a bounded

function, and P is a partition of A . For each subrectangle S of the partition let

$$m_S(f) = \inf\{f(x): x \in S\},$$

$$M_S(f) = \sup\{f(x): x \in S\},$$

and let $v(S)$ be the volume of S [the **volume** of a rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$, and also of $(a_1, b_1) \times \cdots \times (a_n, b_n)$, is defined as $(b_1 - a_1) \cdot \cdots \cdot (b_n - a_n)$]. The **lower** and **upper sums** of f for P are defined by

$$L(f, P) = \sum_S m_S(f) \cdot v(S) \quad \text{and} \quad U(f, P) = \sum_S M_S(f) \cdot v(S).$$

Clearly $L(f, P) \leq U(f, P)$, and an even stronger assertion (3-2) is true.

3-1 Lemma. *Suppose the partition P' refines P (that is, each subrectangle of P' is contained in a subrectangle of P). Then*

$$L(f, P) \leq L(f, P') \quad \text{and} \quad U(f, P') \leq U(f, P).$$

Proof. Each subrectangle S of P is divided into several subrectangles S_1, \dots, S_α of P' , so $v(S) = v(S_1) + \cdots + v(S_\alpha)$. Now $m_S(f) \leq m_{S_i}(f)$, since the values $f(x)$ for $x \in S$ include all values $f(x)$ for $x \in S_i$ (and possibly smaller ones). Thus

$$\begin{aligned} m_S(f) \cdot v(S) &= m_S(f) \cdot v(S_1) + \cdots + m_S(f) \cdot v(S_\alpha) \\ &\leq m_{S_1}(f) \cdot v(S_1) + \cdots + m_{S_\alpha}(f) \cdot v(S_\alpha). \end{aligned}$$

The sum, for all S , of the terms on the left side is $L(f, P)$, while the sum of all the terms on the right side is $L(f, P')$. Hence $L(f, P) \leq L(f, P')$. The proof for upper sums is similar. ■

3-2 Corollary. *If P and P' are any two partitions, then $L(f, P') \leq U(f, P)$.*

Proof. Let P'' be a partition which refines both P and P' . (For example, let $P'' = (P'_1, \dots, P'_n)$, where P'_i is a par-

tion of $[a_i, b_i]$ which refines both P_i and P'_i .) Then

$$L(f, P') \leq L(f, P'') \leq U(f, P'') \leq U(f, P). \quad \blacksquare$$

It follows from Corollary 3-2 that the least upper bound of all lower sums for f is less than or equal to the greatest lower bound of all upper sums for f . A function $f: A \rightarrow \mathbf{R}$ is called **integrable** on the rectangle A if f is bounded and $\sup\{L(f, P)\} = \inf\{U(f, P)\}$. This common number is then denoted $\int_A f$, and called the **integral** of f over A . Often, the notation $\int_A f(x^1, \dots, x^n) dx^1 \cdots dx^n$ is used. If $f: [a, b] \rightarrow \mathbf{R}$, where $a \leq b$, then $\int_a^b f = \int_{[a, b]} f$. A simple but useful criterion for integrability is provided by

3-3 Theorem. *A bounded function $f: A \rightarrow \mathbf{R}$ is integrable if and only if for every $\epsilon > 0$ there is a partition P of A such that $U(f, P) - L(f, P) < \epsilon$.*

Proof. If this condition holds, it is clear that $\sup\{L(f, P)\} = \inf\{U(f, P)\}$ and f is integrable. On the other hand, if f is integrable, so that $\sup\{L(f, P)\} = \inf\{U(f, P)\}$, then for any $\epsilon > 0$ there are partitions P and P' with $U(f, P) - L(f, P') < \epsilon$. If P'' refines both P and P' , it follows from Lemma 3-1 that $U(f, P'') - L(f, P'') \leq U(f, P) - L(f, P') < \epsilon$. \blacksquare

In the following sections we will characterize the integrable functions and discover a method of computing integrals. For the present we consider two functions, one integrable and one not.

1. Let $f: A \rightarrow \mathbf{R}$ be a constant function, $f(x) = c$. Then for any partition P and subrectangle S we have $m_S(f) = M_S(f) = c$, so that $L(f, P) = U(f, P) = \sum_{sc} c \cdot v(S) = c \cdot v(A)$. Hence $\int_A f = c \cdot v(A)$.

2. Let $f: [0, 1] \times [0, 1] \rightarrow \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} 0 & \text{if } x \text{ is rational,} \\ 1 & \text{if } x \text{ is irrational.} \end{cases}$$

If P is a partition, then every subrectangle S will contain points (x, y) with x rational, and also points (x, y) with x

irrational. Hence $m_S(f) = 0$ and $M_S(f) = 1$, so

$$L(f, P) = \sum_S 0 \cdot v(S) = 0$$

and

$$U(f, P) = \sum_S 1 \cdot v(S) = v([0, 1] \times [0, 1]) = 1.$$

Therefore f is not integrable.

Problems. 3-1. Let $f: [0, 1] \times [0, 1] \rightarrow \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} 0 & \text{if } 0 \leq x < \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

Show that f is integrable and $\int_{[0, 1] \times [0, 1]} f = \frac{1}{2}$.

3-2. Let $f: A \rightarrow \mathbf{R}$ be integrable and let $g = f$ except at finitely many points. Show that g is integrable and $\int_A f = \int_A g$.

3-3. Let $f, g: A \rightarrow \mathbf{R}$ be integrable.

(a) For any partition P of A and subrectangle S , show that

$$m_S(f) + m_S(g) \leq m_S(f + g) \quad \text{and} \quad M_S(f + g) \leq M_S(f) + M_S(g)$$

and therefore

$$L(f, P) + L(g, P) \leq L(f + g, P) \quad \text{and} \quad U(f + g, P) \leq U(f, P) + U(g, P).$$

(b) Show that $f + g$ is integrable and $\int_A f + g = \int_A f + \int_A g$.

(c) For any constant c , show that $\int_A cf = c \int_A f$.

3-4. Let $f: A \rightarrow \mathbf{R}$ and let P be a partition of A . Show that f is integrable if and only if for each subrectangle S the function $f|_S$, which consists of f restricted to S , is integrable, and that in this case $\int_A f = \sum_S \int_S f|_S$.

3-5. Let $f, g: A \rightarrow \mathbf{R}$ be integrable and suppose $f \leq g$. Show that $\int_A f \leq \int_A g$.

3-6. If $f: A \rightarrow \mathbf{R}$ is integrable, show that $|f|$ is integrable and $|\int_A f| \leq \int_A |f|$.

3-7. Let $f: [0, 1] \times [0, 1] \rightarrow \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} 0 & x \text{ irrational,} \\ 0 & x \text{ rational, } y \text{ irrational,} \\ 1/q & x \text{ rational, } y = p/q \text{ in lowest terms.} \end{cases}$$

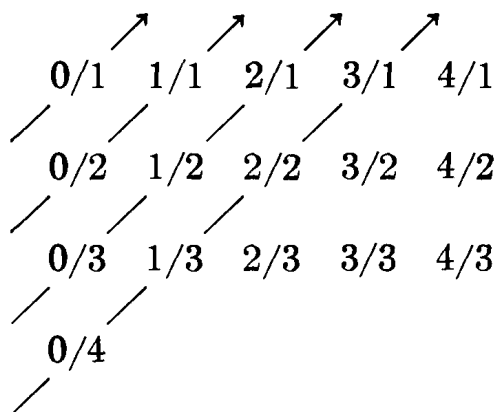
Show that f is integrable and $\int_{[0, 1] \times [0, 1]} f = 0$.

MEASURE ZERO AND CONTENT ZERO

A subset A of \mathbf{R}^n has (n -dimensional) **measure 0** if for every $\varepsilon > 0$ there is a cover $\{U_1, U_2, U_3, \dots\}$ of A by closed rectangles such that $\sum_{i=1}^{\infty} v(U_i) < \varepsilon$. It is obvious (but nevertheless useful to remember) that if A has measure 0 and $B \subset A$, then B has measure 0. The reader may verify that open rectangles may be used instead of closed rectangles in the definition of measure 0.

A set with only finitely many points clearly has measure 0. If A has infinitely many points which can be arranged in a sequence a_1, a_2, a_3, \dots , then A also has measure 0, for if $\varepsilon > 0$, we can choose U_i to be a closed rectangle containing a_i with $v(U_i) < \varepsilon/2^i$. Then $\sum_{i=1}^{\infty} v(U_i) < \sum_{i=1}^{\infty} \varepsilon/2^i = \varepsilon$.

The set of all rational numbers between 0 and 1 is an important and rather surprising example of an infinite set whose members can be arranged in such a sequence. To see that this is so, list the fractions in the following array in the order indicated by the arrows (deleting repetitions and numbers greater than 1):

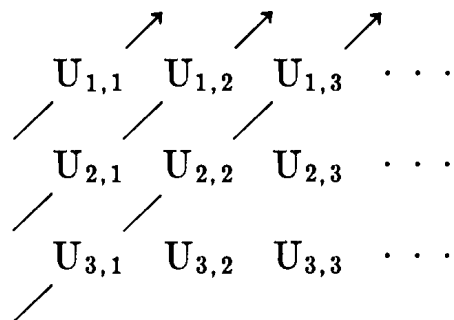


An important generalization of this idea can be given.

3-4 Theorem. *If $A = A_1 \cup A_2 \cup A_3 \cup \dots$ and each A_i has measure 0, then A has measure 0.*

Proof. Let $\varepsilon > 0$. Since A_i has measure 0, there is a cover $\{U_{i,1}, U_{i,2}, U_{i,3}, \dots\}$ of A_i by closed rectangles such that $\sum_{j=1}^{\infty} v(U_{i,j}) < \varepsilon/2^i$. Then the collection of all $U_{i,j}$ is a cover

of A . By considering the array



we see that this collection can be arranged in a sequence V_1, V_2, V_3, \dots . Clearly $\sum_{i=1}^{\infty} v(V_i) < \sum_{i=1}^{\infty} \epsilon/2^i = \epsilon$. ■

A subset A of \mathbf{R}^n has (n -dimensional) **content 0** if for every $\epsilon > 0$ there is a *finite* cover $\{U_1, \dots, U_n\}$ of A by closed rectangles such that $\sum_{i=1}^n v(U_i) < \epsilon$. If A has content 0, then A clearly has measure 0. Again, open rectangles could be used instead of closed rectangles in the definition.

3-5 Theorem. *If $a < b$, then $[a, b] \subset \mathbf{R}$ does not have content 0. In fact, if $\{U_1, \dots, U_n\}$ is a finite cover of $[a, b]$ by closed intervals, then $\sum_{i=1}^n v(U_i) \geq b - a$.*

Proof. Clearly we can assume that each $U_i \subset [a, b]$. Let $a = t_0 < t_1 < \dots < t_k = b$ be all endpoints of all U_i . Then each $v(U_i)$ is the sum of certain $t_j - t_{j-1}$. Moreover, each $[t_{j-1}, t_j]$ lies in at least one U_i (namely, any one which contains an interior point of $[t_{j-1}, t_j]$), so $\sum_{i=1}^n v(U_i) \geq \sum_{j=1}^k (t_j - t_{j-1}) = b - a$. ■

If $a < b$, it is also true that $[a, b]$ does not have measure 0. This follows from

3-6 Theorem. *If A is compact and has measure 0, then A has content 0.*

Proof. Let $\epsilon > 0$. Since A has measure 0, there is a cover $\{U_1, U_2, \dots\}$ of A by open rectangles such that $\sum_{i=1}^{\infty} v(U_i)$

$< \varepsilon$. Since A is compact, a finite number U_1, \dots, U_n of the U_i also cover A and surely $\sum_{i=1}^n v(U_i) < \varepsilon$. ■

The conclusion of Theorem 3-6 is false if A is not compact. For example, let A be the set of rational numbers between 0 and 1; then A has measure 0. Suppose, however, that $\{[a_1, b_1], \dots, [a_n, b_n]\}$ covers A . Then A is contained in the closed set $[a_1, b_1] \cup \dots \cup [a_n, b_n]$, and therefore $[0, 1] \subset [a_1, b_1] \cup \dots \cup [a_n, b_n]$. It follows from Theorem 3-5 that $\sum_{i=1}^n (b_i - a_i) \geq 1$ for any such cover, and consequently A does not have content 0.

Problems. 3-8. Prove that $[a_1, b_1] \times \dots \times [a_n, b_n]$ does not have content 0 if $a_i < b_i$ for each i .

3-9. (a) Show that an unbounded set cannot have content 0.

(b) Give an example of a closed set of measure 0 which does not have content 0.

3-10. (a) If C is a set of content 0, show that the boundary of C has content 0.

(b) Give an example of a bounded set C of measure 0 such that the boundary of C does not have measure 0.

3-11. Let A be the set of Problem 1-18. If $\sum_{i=1}^{\infty} (b_i - a_i) < 1$, show that the boundary of A does not have measure 0.

3-12. Let $f: [a, b] \rightarrow \mathbf{R}$ be an increasing function. Show that $\{x: f \text{ is discontinuous at } x\}$ has measure 0. *Hint:* Use Problem 1-30 to show that $\{x: o(f, x) > 1/n\}$ is finite, for each integer n .

3-13.* (a) Show that the collection of all rectangles $[a_1, b_1] \times \dots \times [a_n, b_n]$ with all a_i and b_i rational can be arranged in a sequence.

(b) If $A \subset \mathbf{R}^n$ is any set and \mathcal{O} is an open cover of A , show that there is a sequence U_1, U_2, U_3, \dots of members of \mathcal{O} which also cover A . *Hint:* For each $x \in A$ there is a rectangle $B = [a_1, b_1] \times \dots \times [a_n, b_n]$ with all a_i and b_i rational such that $x \in B \subset U$ for some $U \in \mathcal{O}$.

INTEGRABLE FUNCTIONS

Recall that $o(f, x)$ denotes the oscillation of f at x .

3-7 Lemma. Let A be a closed rectangle and let $f: A \rightarrow \mathbf{R}$ be a bounded function such that $o(f, x) < \varepsilon$ for all $x \in A$. Then there is a partition P of A with $U(f, P) - L(f, P) < \varepsilon \cdot v(A)$.

Proof. For each $x \in A$ there is a closed rectangle U_x , containing x in its interior, such that $M_{U_x}(f) - m_{U_x}(f) < \epsilon$. Since A is compact, a finite number U_{x_1}, \dots, U_{x_n} of the sets U_x cover A . Let P be a partition for A such that each subrectangle S of P is contained in some U_{x_i} . Then $M_S(f) - m_S(f) < \epsilon$ for each subrectangle S of P , so that $U(f, P) - L(f, P) = \sum_S [M_S(f) - m_S(f)] \cdot v(S) < \epsilon \cdot v(A)$. ■

3-8 Theorem. Let A be a closed rectangle and $f: A \rightarrow \mathbf{R}$ a bounded function. Let $B = \{x: f \text{ is not continuous at } x\}$. Then f is integrable if and only if B is a set of measure 0.

Proof. Suppose first that B has measure 0. Let $\epsilon > 0$ and let $B_\epsilon = \{x: o(f, x) \geq \epsilon\}$. Then $B_\epsilon \subset B$, so that B_ϵ has measure 0. Since (Theorem 1-11) B_ϵ is compact, B_ϵ has content 0. Thus there is a finite collection U_1, \dots, U_n of closed rectangles, whose interiors cover B_ϵ , such that $\sum_{i=1}^n v(U_i) < \epsilon$. Let P be a partition of A such that every subrectangle S of P is in one of two groups (see Figure 3-1):

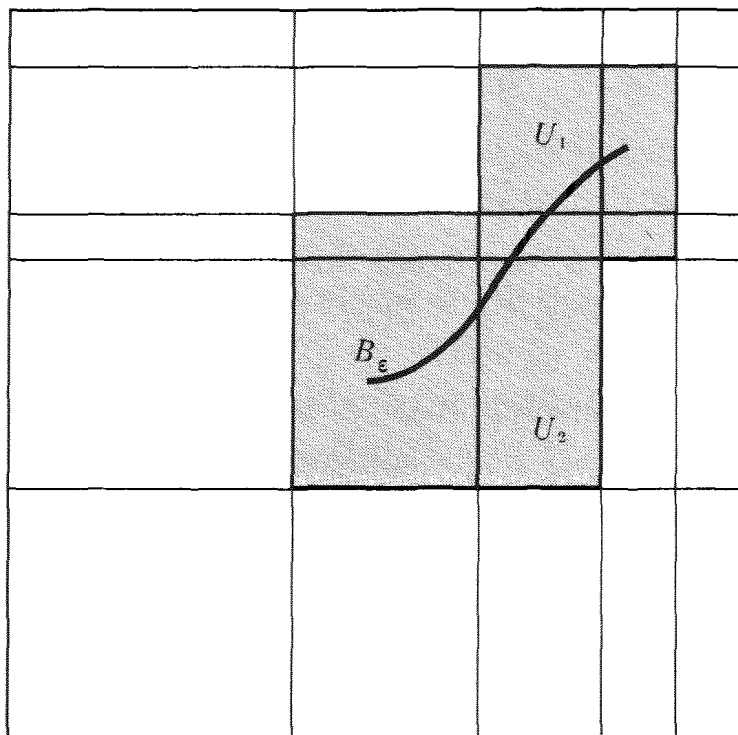


FIGURE 3-1. The shaded rectangles are in \mathcal{S}_1 .

- (1) \mathfrak{S}_1 , which consists of subrectangles S , such that $S \subset U_i$ for some i .
- (2) \mathfrak{S}_2 , which consists of subrectangles S with $S \cap B_\varepsilon = \emptyset$.

Let $|f(x)| < M$ for $x \in A$. Then $M_S(f) - m_S(f) < 2M$ for every S . Therefore

$$\sum_{S \in \mathfrak{S}_1} [M_S(f) - m_S(f)] \cdot v(S) < 2M \sum_{i=1}^n v(U_i) < 2M\varepsilon.$$

Now, if $S \in \mathfrak{S}_2$, then $o(f, x) < \varepsilon$ for $x \in S$. Lemma 3-7 implies that there is a refinement P' of P such that

$$\sum_{S' \subset S} [M_{S'}(f) - m_{S'}(f)] \cdot v(S') < \varepsilon \cdot v(S)$$

for $S \in \mathfrak{S}_2$. Then

$$\begin{aligned} U(f, P') - L(f, P') &= \sum_{S' \subset S \in \mathfrak{S}_1} [M_{S'}(f) - m_{S'}(f)] \cdot v(S') \\ &\quad + \sum_{S' \subset S \in \mathfrak{S}_2} [M_{S'}(f) - m_{S'}(f)] \cdot v(S') \\ &< 2M\varepsilon + \sum_{S \in \mathfrak{S}_2} \varepsilon \cdot v(S) \\ &\leq 2M\varepsilon + \varepsilon \cdot v(A). \end{aligned}$$

Since M and $v(A)$ are fixed, this shows that we can find a partition P' with $U(f, P') - L(f, P')$ as small as desired. Thus f is integrable.

Suppose, conversely, that f is integrable. Since $B = B_1 \cup B_{\frac{1}{2}} \cup B_{\frac{1}{3}} \cup \cdots$, it suffices (Theorem 3-4) to prove that each $B_{1/n}$ has measure 0. In fact we will show that each $B_{1/n}$ has content 0 (since $B_{1/n}$ is compact, this is actually equivalent).

If $\varepsilon > 0$, let P be a partition of A such that $U(f, P) - L(f, P) < \varepsilon/n$. Let \mathfrak{S} be the collection of subrectangles S of P which intersect $B_{1/n}$. Then \mathfrak{S} is a cover of $B_{1/n}$. Now if

$S \in \mathfrak{S}$, then $M_S(f) - m_S(f) \geq 1/n$. Thus

$$\begin{aligned} \frac{1}{n} \cdot \sum_{S \in \mathfrak{S}} v(S) &\leq \sum_{S \in \mathfrak{S}} [M_S(f) - m_S(f)] \cdot v(S) \\ &\leq \sum_S [M_S(f) - m_S(f)] \cdot v(S) \\ &< \frac{\varepsilon}{n}, \end{aligned}$$

and consequently $\sum_{S \in \mathfrak{S}} v(S) < \varepsilon$. ■

We have thus far dealt only with the integrals of functions over rectangles. Integrals over other sets are easily reduced to this type. If $C \subset \mathbf{R}^n$, the **characteristic function** χ_C of C is defined by

$$\chi_C(x) = \begin{cases} 0 & x \notin C, \\ 1 & x \in C. \end{cases}$$

If $C \subset A$ for some closed rectangle A and $f: A \rightarrow \mathbf{R}$ is bounded, then $\int_C f$ is defined as $\int_A f \cdot \chi_C$, provided $f \cdot \chi_C$ is integrable. This certainly occurs (Problem 3-14) if f and χ_C are integrable.

3-9 Theorem. *The function $\chi_C: A \rightarrow \mathbf{R}$ is integrable if and only if the boundary of C has measure 0 (and hence content 0).*

Proof. If x is in the interior of C , then there is an open rectangle U with $x \in U \subset C$. Thus $\chi_C = 1$ on U and χ_C is clearly continuous at x . Similarly, if x is in the exterior of C , there is an open rectangle U with $x \in U \subset \mathbf{R}^n - C$. Hence $\chi_C = 0$ on U and χ_C is continuous at x . Finally, if x is in the boundary of C , then for every open rectangle U containing x , there is $y_1 \in U \cap C$, so that $\chi_C(y_1) = 1$ and there is $y_2 \in U \cap (\mathbf{R}^n - C)$, so that $\chi_C(y_2) = 0$. Hence χ_C is not continuous at x . Thus $\{x: \chi_C \text{ is not continuous at } x\} = \text{boundary } C$, and the result follows from Theorem 3-8. ■

A bounded set C whose boundary has measure 0 is called **Jordan-measurable**. The integral $\int_C 1$ is called the (n -dimensional) **content** of C , or the (n -dimensional) **volume** of C . Naturally one-dimensional volume is often called **length**, and two-dimensional volume, **area**.

Problem 3-11 shows that even an open set C may not be Jordan-measurable, so that $\int_C f$ is not necessarily defined even if C is open and f is continuous. This unhappy state of affairs will be rectified soon.

Problems. 3-14. Show that if $f, g: A \rightarrow \mathbf{R}$ are integrable, so is $f \cdot g$.

3-15. Show that if C has content 0, then $C \subset A$ for some closed rectangle A and C is Jordan-measurable and $\int_A \chi_C = 0$.

3-16. Give an example of a bounded set C of measure 0 such that $\int_A \chi_C$ does not exist.

3-17. If C is a bounded set of measure 0 and $\int_A \chi_C$ exists, show that $\int_A \chi_C = 0$. *Hint:* Show that $L(f, P) = 0$ for all partitions P . Use Problem 3-8.

3-18. If $f: A \rightarrow \mathbf{R}$ is non-negative and $\int_A f = 0$, show that $\{x: f(x) \neq 0\}$ has measure 0. *Hint:* Prove that $\{x: f(x) > 1/n\}$ has content 0.

3-19. Let U be the open set of Problem 3-11. Show that if $f = \chi_U$ except on a set of measure 0, then f is not integrable on $[0, 1]$.

3-20. Show that an increasing function $f: [a, b] \rightarrow \mathbf{R}$ is integrable on $[a, b]$.

3-21. If A is a closed rectangle, show that $C \subset A$ is Jordan-measurable if and only if for every $\varepsilon > 0$ there is a partition P of A such that $\sum_{S \in \mathcal{S}_1} v(S) - \sum_{S \in \mathcal{S}_2} v(S) < \varepsilon$, where \mathcal{S}_1 consists of all subrectangles intersecting C and \mathcal{S}_2 all subrectangles contained in C .

3-22.* If A is a Jordan-measurable set and $\varepsilon > 0$, show that there is a compact Jordan-measurable set $C \subset A$ such that $\int_{A-C} 1 < \varepsilon$.

FUBINI'S THEOREM

The problem of calculating integrals is solved, in some sense, by Theorem 3-10, which reduces the computation of integrals over a closed rectangle in \mathbf{R}^n , $n > 1$, to the computation of integrals over closed intervals in \mathbf{R} . Of sufficient importance to deserve a special designation, this theorem is usually referred to as Fubini's theorem, although it is more or less a

special case of a theorem proved by Fubini long after Theorem 3-10 was known.

The idea behind the theorem is best illustrated (Figure 3-2) for a positive continuous function $f: [a,b] \times [c,d] \rightarrow \mathbf{R}$. Let t_0, \dots, t_n be a partition of $[a,b]$ and divide $[a,b] \times [c,d]$ into n strips by means of the line segments $\{t_i\} \times [c,d]$. If g_x is defined by $g_x(y) = f(x,y)$, then the area of the region under the graph of f and above $\{x\} \times [c,d]$ is

$$\int_c^d g_x = \int_c^d f(x,y)dy.$$

The volume of the region under the graph of f and above $[t_{i-1}, t_i] \times [c,d]$ is therefore approximately equal to $(t_i - t_{i-1}) \cdot \int_c^d f(x,y)dy$, for any $x \in [t_{i-1}, t_i]$. Thus

$$\int_{[a,b] \times [c,d]} f = \sum_{i=1}^n \int_{[t_{i-1}, t_i] \times [c,d]} f$$

is approximately $\sum_{i=1}^n (t_i - t_{i-1}) \cdot \int_c^d f(x_i, y)dy$, with x_i in

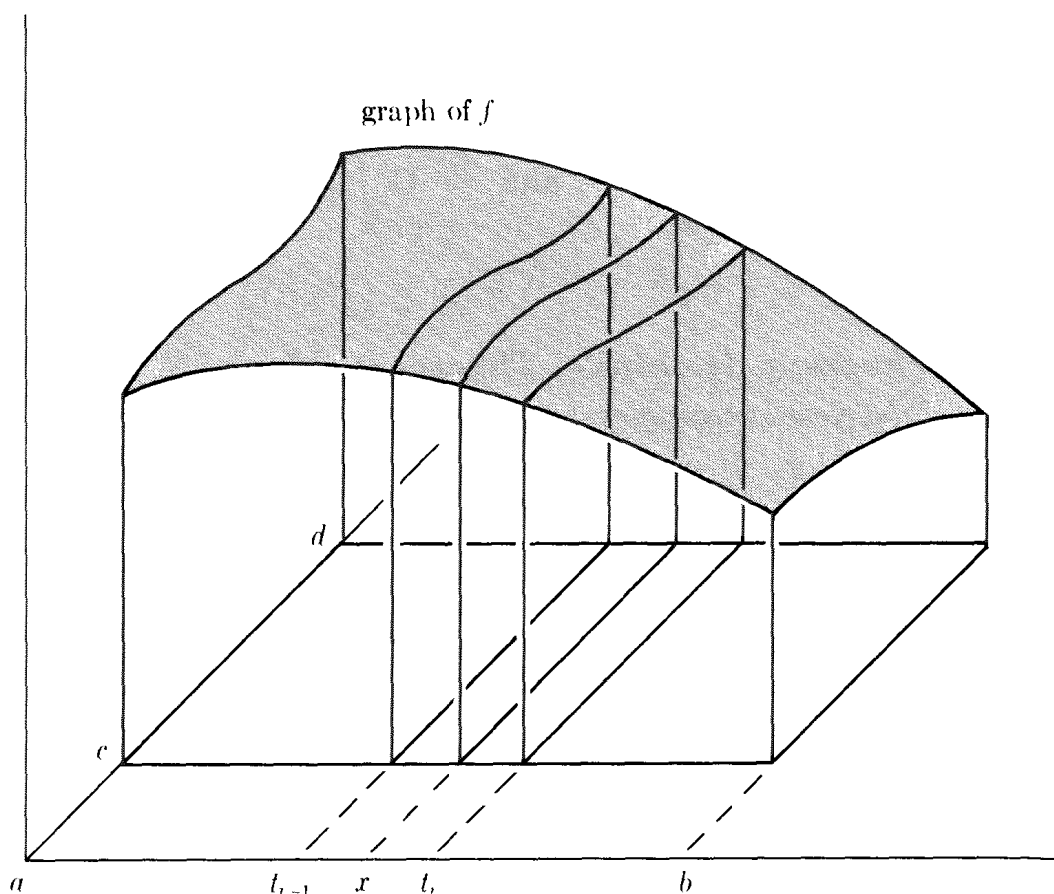


FIGURE 3-2

$[t_{i-1}, t_i]$. On the other hand, sums similar to these appear in the definition of $\int_a^b (\int_c^d f(x, y) dy) dx$. Thus, if h is defined by $h(x) = \int_c^d g_x = \int_c^d f(x, y) dy$, it is reasonable to hope that h is integrable on $[a, b]$ and that

$$\int_{[a,b] \times [c,d]} f = \int_a^b h = \int_a^b \left(\int_c^d f(x, y) dy \right) dx.$$

This will indeed turn out to be true when f is continuous, but in the general case difficulties arise. Suppose, for example, that the set of discontinuities of f is $\{x_0\} \times [c, d]$ for some $x_0 \in [a, b]$. Then f is integrable on $[a, b] \times [c, d]$ but $h(x_0) = \int_c^d f(x_0, y) dy$ may not even be defined. The statement of Fubini's theorem therefore looks a little strange, and will be followed by remarks about various special cases where simpler statements are possible.

We will need one bit of terminology. If $f: A \rightarrow \mathbf{R}$ is a bounded function on a closed rectangle, then, whether or not f is integrable, the least upper bound of all lower sums, and the greatest lower bound of all upper sums, both exist. They are called the **lower** and **upper integrals** of f on A , and denoted

$$\mathbf{L} \int_A f \quad \text{and} \quad \mathbf{U} \int_A f,$$

respectively.

3-10 Theorem (Fubini's Theorem). *Let $A \subset \mathbf{R}^n$ and $B \subset \mathbf{R}^m$ be closed rectangles, and let $f: A \times B \rightarrow \mathbf{R}$ be integrable. For $x \in A$ let $g_x: B \rightarrow \mathbf{R}$ be defined by $g_x(y) = f(x, y)$ and let*

$$\begin{aligned} \mathfrak{L}(x) &= \mathbf{L} \int_B g_x = \mathbf{L} \int_B f(x, y) dy, \\ \mathfrak{U}(x) &= \mathbf{U} \int_B g_x = \mathbf{U} \int_B f(x, y) dy. \end{aligned}$$

Then \mathfrak{L} and \mathfrak{U} are integrable on A and

$$\begin{aligned} \int_{A \times B} f &= \int_A \mathfrak{L} = \int_A \left(\mathbf{L} \int_B f(x, y) dy \right) dx, \\ \int_{A \times B} f &= \int_A \mathfrak{U} = \int_A \left(\mathbf{U} \int_B f(x, y) dy \right) dx. \end{aligned}$$

(The integrals on the right side are called **iterated integrals** for f .)

Proof. Let P_A be a partition of A and P_B a partition of B . Together they give a partition P of $A \times B$ for which any subrectangle S is of the form $S_A \times S_B$, where S_A is a subrectangle of the partition P_A , and S_B is a subrectangle of the partition P_B . Thus

$$\begin{aligned} L(f, P) &= \sum_S m_S(f) \cdot v(S) = \sum_{S_A, S_B} m_{S_A \times S_B}(f) \cdot v(S_A \times S_B) \\ &= \sum_{S_A} \left(\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \right) \cdot v(S_A). \end{aligned}$$

Now, if $x \in S_A$, then clearly $m_{S_A \times S_B}(f) \leq m_{S_B}(g_x)$. Consequently, for $x \in S_A$ we have

$$\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \leq \sum_{S_B} m_{S_B}(g_x) \cdot v(S_B) \leq \mathbf{L} \int_B g_x = \mathfrak{L}(x).$$

Therefore

$$\sum_{S_A} \left(\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \right) \cdot v(S_A) \leq L(\mathfrak{L}, P_A).$$

We thus obtain

$$L(f, P) \leq L(\mathfrak{L}, P_A) \leq U(\mathfrak{L}, P_A) \leq U(\mathfrak{U}, P_A) \leq U(f, P),$$

where the proof of the last inequality is entirely analogous to the proof of the first. Since f is integrable, $\sup\{L(f, P)\} = \inf\{U(f, P)\} = \int_{A \times B} f$. Hence

$$\sup\{L(\mathfrak{L}, P_A)\} = \inf\{U(\mathfrak{L}, P_A)\} = \int_{A \times B} f.$$

In other words, \mathfrak{L} is integrable on A and $\int_{A \times B} f = \int_A \mathfrak{L}$. The assertion for \mathfrak{U} follows similarly from the inequalities

$$L(f, P) \leq L(\mathfrak{L}, P_A) \leq L(\mathfrak{U}, P_A) \leq U(\mathfrak{U}, P_A) \leq U(f, P). \quad \blacksquare$$

Remarks. 1. A similar proof shows that

$$\int_{A \times B} f = \int_B \left(\mathbf{L} \int_A f(x, y) dx \right) dy = \int_B \left(\mathbf{U} \int_A f(x, y) dx \right) dy.$$

These integrals are called *iterated integrals* for f in the reverse order from those of the theorem. As several problems show, the possibility of interchanging the orders of iterated integrals has many consequences.

2. In practice it is often the case that each g_x is integrable, so that $\int_{A \times B} f = \int_A (\int_B f(x, y) dy) dx$. This certainly occurs if f is continuous.

3. The worst irregularity commonly encountered is that g_x is not integrable for a finite number of $x \in A$. In this case $\mathcal{L}(x) = \int_B f(x, y) dy$ for all but these finitely many x . Since $\int_A \mathcal{L}$ remains unchanged if \mathcal{L} is redefined at a finite number of points, we can still write $\int_{A \times B} f = \int_A (\int_B f(x, y) dy) dx$, provided that $\int_B f(x, y) dy$ is defined arbitrarily, say as 0, when it does not exist.

4. There are cases when this will not work and Theorem 3-10 must be used as stated. Let $f: [0, 1] \times [0, 1] \rightarrow \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} 1 & \text{if } x \text{ is irrational,} \\ 1 & \text{if } x \text{ is rational and } y \text{ is irrational,} \\ 1 - 1/q & \text{if } x = p/q \text{ in lowest terms and } y \text{ is} \\ & \text{rational.} \end{cases}$$

Then f is integrable and $\int_{[0, 1] \times [0, 1]} f = 1$. Now $\int_0^1 f(x, y) dy = 1$ if x is irrational, and does not exist if x is rational. Therefore h is not integrable if $h(x) = \int_0^1 f(x, y) dy$ is set equal to 0 when the integral does not exist.

5. If $A = [a_1, b_1] \times \cdots \times [a_n, b_n]$ and $f: A \rightarrow \mathbf{R}$ is sufficiently nice, we can apply Fubini's theorem repeatedly to obtain

$$\int_A f = \int_{a_n}^{b_n} \left(\cdots \left(\int_{a_1}^{b_1} f(x^1, \dots, x^n) dx^1 \right) \cdots \right) dx^n.$$

6. If $C \subset A \times B$, Fubini's theorem can be used to evaluate $\int_C f$, since this is by definition $\int_{A \times B} \chi_C f$. Suppose, for example, that

$$C = [-1, 1] \times [-1, 1] - \{(x, y): |(x, y)| < 1\}.$$

Then

$$\int_C f = \int_{-1}^1 \left(\int_{-1}^1 f(x, y) \cdot \chi_C(x, y) dy \right) dx.$$

Now

$$\chi_C(x,y) = \begin{cases} 1 & \text{if } y > \sqrt{1-x^2} \text{ or } y < -\sqrt{1-x^2}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$\int_{-1}^1 f(x,y) \cdot \chi_C(x,y) dy = \int_{-1}^{-\sqrt{1-x^2}} f(x,y) dy + \int_{\sqrt{1-x^2}}^1 f(x,y) dy.$$

In general, if $C \subset A \times B$, the main difficulty in deriving expressions for $\int_C f$ will be determining $C \cap (\{x\} \times B)$ for $x \in A$. If $C \cap (A \times \{y\})$ for $y \in B$ is easier to determine, one should use the iterated integral

$$\int_C f = \int_B \left(\int_A f(x,y) \cdot \chi_C(x,y) dx \right) dy.$$

Problems. 3-23. Let $C \subset A \times B$ be a set of content 0. Let $A' \subset A$ be the set of all $x \in A$ such that $\{y \in B: (x,y) \in C\}$ is not of content 0. Show that A' is a set of measure 0. *Hint:* χ_C is integrable and $\int_{A \times B} \chi_C = \int_A \mathfrak{U} = \int_A \mathfrak{L}$, so $\int_A \mathfrak{U} - \mathfrak{L} = 0$.

3-24. Let $C \subset [0,1] \times [0,1]$ be the union of all $\{p/q\} \times [0, 1/q]$, where p/q is a rational number in $[0,1]$ written in lowest terms. Use C to show that the word “measure” in Problem 3-23 cannot be replaced by “content.”

3-25. Use induction on n to show that $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is not a set of measure 0 (or content 0) if $a_i < b_i$ for each i .

3-26. Let $f: [a,b] \rightarrow \mathbf{R}$ be integrable and non-negative and let $A_f = \{(x,y): a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}$. Show that A_f is Jordan-measurable and has area $\int_a^b f$.

3-27. If $f: [a,b] \times [a,b] \rightarrow \mathbf{R}$ is continuous, show that

$$\int_a^b \int_a^y f(x,y) dx dy = \int_a^b \int_x^b f(x,y) dy dx.$$

Hint: Compute $\int_C f$ in two different ways for a suitable set $C \subset [a,b] \times [a,b]$.

3-28.* Use Fubini's theorem to give an easy proof that $D_{1,2}f = D_{2,1}f$ if these are continuous. *Hint:* If $D_{1,2}f(a) - D_{2,1}f(a) > 0$, there is a rectangle A containing a such that $D_{1,2}f - D_{2,1}f > 0$ on A .

3-29. Use Fubini's theorem to derive an expression for the volume of a set of \mathbf{R}^3 obtained by revolving a Jordan-measurable set in the yz -plane about the z -axis.

3-30. Let C be the set in Problem 1-17. Show that

$$\int_{[0,1]} \left(\int_{[0,1]} \chi_C(x,y) dx \right) dy = \int_{[0,1]} \left(\int_{[0,1]} \chi_C(y,x) dy \right) dx = 0$$

but that $\int_{[0,1] \times [0,1]} \chi_C$ does not exist.

3-31. If $A = [a_1, b_1] \times \cdots \times [a_n, b_n]$ and $f: A \rightarrow \mathbf{R}$ is continuous, define $F: A \rightarrow \mathbf{R}$ by

$$F(x) = \int_{[a_1, x_1] \times \cdots \times [a_n, x_n]} f.$$

What is $D_i F(x)$, for x in the interior of A ?

3-32.* Let $f: [a, b] \times [c, d] \rightarrow \mathbf{R}$ be continuous and suppose $D_2 f$ is continuous. Define $F(y) = \int_a^b f(x, y) dx$. Prove *Leibnitz's rule*: $F'(y) = \int_a^b D_2 f(x, y) dx$. *Hint*: $F(y) = \int_a^b f(x, y) dx = \int_a^b \left(\int_c^y D_2 f(x, y) dy + f(x, c) \right) dx$. (The proof will show that continuity of $D_2 f$ may be replaced by considerably weaker hypotheses.)

3-33. If $f: [a, b] \times [c, d] \rightarrow \mathbf{R}$ is continuous and $D_2 f$ is continuous, define $F(x, y) = \int_a^x f(t, y) dt$.

(a) Find $D_1 F$ and $D_2 F$.

(b) If $G(x) = \int_a^{g(x)} f(t, x) dt$, find $G'(x)$.

3-34.* Let $g_1, g_2: \mathbf{R}^2 \rightarrow \mathbf{R}$ be continuously differentiable and suppose $D_1 g_2 = D_2 g_1$. As in Problem 2-21, let

$$f(x, y) = \int_0^x g_1(t, 0) dt + \int_0^y g_2(x, t) dt.$$

Show that $D_1 f(x, y) = g_1(x, y)$.

3-35.* (a) Let $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$ be a linear transformation of one of the following types:

$$\begin{cases} g(e_i) = e_i & i \neq j \\ g(e_j) = a e_j \end{cases}$$

$$\begin{cases} g(e_i) = e_i & i \neq j \\ g(e_j) = e_j + e_k \end{cases}$$

$$\begin{cases} g(e_k) = e_k & k \neq i, j \\ g(e_i) = e_j \\ g(e_j) = e_i. \end{cases}$$

If U is a rectangle, show that the volume of $g(U)$ is $|\det g| \cdot v(U)$.

(b) Prove that $|\det g| \cdot v(U)$ is the volume of $g(U)$ for any linear transformation $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$. *Hint*: If $\det g \neq 0$, then g is the composition of linear transformations of the type considered in (a).

3-36. (Cavalieri's principle). Let A and B be Jordan-measurable subsets of \mathbf{R}^3 . Let $A_c = \{(x, y): (x, y, c) \in A\}$ and define B_c similarly. Suppose each A_c and B_c are Jordan-measurable and have the same area. Show that A and B have the same volume.

PARTITIONS OF UNITY

In this section we introduce a tool of extreme importance in the theory of integration.

3-11 Theorem. *Let $A \subset \mathbf{R}^n$ and let \mathcal{O} be an open cover of A . Then there is a collection Φ of C^∞ functions φ defined in an open set containing A , with the following properties:*

- (1) *For each $x \in A$ we have $0 \leq \varphi(x) \leq 1$.*
- (2) *For each $x \in A$ there is an open set V containing x such that all but finitely many $\varphi \in \Phi$ are 0 on V .*
- (3) *For each $x \in A$ we have $\sum_{\varphi \in \Phi} \varphi(x) = 1$ (by (2) for each x this sum is finite in some open set containing x).*
- (4) *For each $\varphi \in \Phi$ there is an open set U in \mathcal{O} such that $\varphi = 0$ outside of some closed set contained in U .*

(A collection Φ satisfying (1) to (3) is called a C^∞ **partition of unity** for A . If Φ also satisfies (4), it is said to be **subordinate** to the cover \mathcal{O} . In this chapter we will only use continuity of the functions φ .)

Proof. *Case 1. A is compact.*

Then a finite number U_1, \dots, U_n of open sets in \mathcal{O} cover A . It clearly suffices to construct a partition of unity subordinate to the cover $\{U_1, \dots, U_n\}$. We will first find compact sets $D_i \subset U_i$ whose interiors cover A . The sets D_i are constructed inductively as follows. Suppose that D_1, \dots, D_k have been chosen so that $\{\text{interior } D_1, \dots, \text{interior } D_k, U_{k+1}, \dots, U_n\}$ covers A . Let

$$C_{k+1} = A - (\text{int } D_1 \cup \dots \cup \text{int } D_k \cup U_{k+2} \cup \dots \cup U_n).$$

Then $C_{k+1} \subset U_{k+1}$ is compact. Hence (Problem 1-22) we can find a compact set D_{k+1} such that

$$C_{k+1} \subset \text{interior } D_{k+1} \quad \text{and} \quad D_{k+1} \subset U_{k+1}.$$

Having constructed the sets D_1, \dots, D_n , let ψ_i be a non-negative C^∞ function which is positive on D_i and 0 outside of some closed set contained in U_i (Problem 2-26). Since

$\{D_1, \dots, D_n\}$ covers A , we have $\psi_1(x) + \dots + \psi_n(x) > 0$ for all x in some open set U containing A . On U we can define

$$\varphi_i(x) = \frac{\psi_i(x)}{\psi_1(x) + \dots + \psi_n(x)}.$$

If $f: U \rightarrow [0,1]$ is a C^∞ function which is 1 on A and 0 outside of some closed set in U , then $\Phi = \{f \cdot \varphi_1, \dots, f \cdot \varphi_n\}$ is the desired partition of unity.

Case 2. $A = A_1 \cup A_2 \cup A_3 \cup \dots$, where each A_i is compact and $A_i \subset \text{interior } A_{i+1}$.

For each i let Θ_i consist of all $U \cap (\text{interior } A_{i+1} - A_{i-2})$ for U in \mathcal{O} . Then Θ_i is an open cover of the compact set $B_i = A_i - \text{interior } A_{i-1}$. By case 1 there is a partition of unity Φ_i for B_i , subordinate to Θ_i . For each $x \in A$ the sum

$$\sigma(x) = \sum_{\varphi \in \Phi_i, \text{ all } i} \varphi(x)$$

is a finite sum in some open set containing x , since if $x \in A_i$ we have $\varphi(x) = 0$ for $\varphi \in \Phi_j$ with $j \geq i+2$. For each φ in each Φ_i , define $\varphi'(x) = \varphi(x)/\sigma(x)$. The collection of all φ' is the desired partition of unity.

Case 3. A is open.

Let $A_i =$

$\{x \in A: |x| \leq i \text{ and distance from } x \text{ to boundary } A \geq 1/i\},$

and apply case 2.

Case 4. A is arbitrary.

Let B be the union of all U in \mathcal{O} . By case 3 there is a partition of unity for B ; this is also a partition of unity for A . ■

An important consequence of condition (2) of the theorem should be noted. Let $C \subset A$ be compact. For each $x \in C$ there is an open set V_x containing x such that only finitely many $\varphi \in \Phi$ are not 0 on V_x . Since C is compact, finitely many such V_x cover C . Thus only finitely many $\varphi \in \Phi$ are not 0 on C .

One important application of partitions of unity will illustrate their main role—piecing together results obtained locally.

An open cover \mathcal{O} of an open set $A \subset \mathbf{R}^n$ is **admissible** if each $U \in \mathcal{O}$ is contained in A . If Φ is subordinate to \mathcal{O} , $f: A \rightarrow \mathbf{R}$ is bounded in some open set around each point of A , and $\{x: f \text{ is discontinuous at } x\}$ has measure 0, then each $\int_A \varphi \cdot |f|$ exists. We define f to be **integrable** (in the extended sense) if $\sum_{\varphi \in \Phi} \int_A \varphi \cdot |f|$ converges (the proof of Theorem 3-11 shows that the φ 's may be arranged in a sequence). This implies convergence of $\sum_{\varphi \in \Phi} \left| \int_A \varphi \cdot f \right|$, and hence absolute convergence of $\sum_{\varphi \in \Phi} \int_A \varphi \cdot f$, which we define to be $\int_A f$. These definitions do not depend on \mathcal{O} or Φ (but see Problem 3-38).

3-12 Theorem.

- (1) If Ψ is another partition of unity, subordinate to an admissible cover \mathcal{O}' of A , then $\sum_{\psi \in \Psi} \int_A \psi \cdot |f|$ also converges, and

$$\sum_{\varphi \in \Phi} \int_A \varphi \cdot f = \sum_{\psi \in \Psi} \int_A \psi \cdot f.$$

- (2) If A and f are bounded, then f is integrable in the extended sense.
 (3) If A is Jordan-measurable and f is bounded, then this definition of $\int_A f$ agrees with the old one.

Proof

- (1) Since $\varphi \cdot f = 0$ except on some compact set C , and there are only finitely many ψ which are non-zero on C , we can write

$$\sum_{\varphi \in \Phi} \int_A \varphi \cdot f = \sum_{\varphi \in \Phi} \int_A \sum_{\psi \in \Psi} \psi \cdot \varphi \cdot f = \sum_{\varphi \in \Phi} \sum_{\psi \in \Psi} \int_A \psi \cdot \varphi \cdot f.$$

This result, applied to $|f|$, shows the convergence of $\sum_{\varphi \in \Phi} \sum_{\psi \in \Psi} \int_A \psi \cdot \varphi \cdot |f|$, and hence of $\sum_{\varphi \in \Phi} \sum_{\psi \in \Psi} \left| \int_A \psi \cdot \varphi \cdot f \right|$. This absolute convergence justifies interchanging the order of summation in the above equation; the resulting double sum clearly equals $\sum_{\psi \in \Psi} \int_A \psi \cdot f$. Finally, this result applied to $|f|$ proves convergence of $\sum_{\psi \in \Psi} \int_A \psi \cdot |f|$.

- (2) If A is contained in the closed rectangle B and $|f(x)| \leq M$ for $x \in A$, and $F \subset \Phi$ is finite, then

$$\sum_{\varphi \in F} \int_A \varphi \cdot |f| \leq \sum_{\varphi \in F} M \int_A \varphi = M \int_A \sum_{\varphi \in F} \varphi \leq Mv(B),$$

since $\sum_{\varphi \in F} \varphi \leq 1$ on A .

- (3) If $\varepsilon > 0$ there is (Problem 3-22) a compact Jordan-measurable $C \subset A$ such that $\int_{A-C} 1 < \varepsilon$. There are only finitely many $\varphi \in \Phi$ which are non-zero on C . If $F \subset \Phi$ is any finite collection which includes these, and $\int_A f$ has its old meaning, then

$$\begin{aligned} \left| \int_A f - \sum_{\varphi \in F} \int_A \varphi \cdot f \right| &\leq \int_A \left| f - \sum_{\varphi \in F} \varphi \cdot f \right| \\ &\leq M \int_A \left(1 - \sum_{\varphi \in F} \varphi \right) \\ &= M \int_A \sum_{\varphi \in \Phi - F} \varphi \leq M \int_{A-C} 1 \leq M\varepsilon. \quad \blacksquare \end{aligned}$$

Problems. 3-37. (a) Suppose that $f: (0,1) \rightarrow \mathbf{R}$ is a non-negative continuous function. Show that $\int_{(0,1)} f$ exists if and only if $\lim_{\varepsilon \rightarrow 0} \int_{\varepsilon}^{1-\varepsilon} f$ exists.

(b) Let $A_n = [1 - 1/2^n, 1 - 1/2^{n+1}]$. Suppose that $f: (0,1) \rightarrow \mathbf{R}$ satisfies $\int_{A_n} f = (-1)^n/n$ and $f(x) = 0$ for $x \notin \text{any } A_n$. Show that $\int_{(0,1)} f$ does not exist, but $\lim_{\varepsilon \rightarrow 0} \int_{(\varepsilon, 1-\varepsilon)} f = \log 2$.

- 3-38. Let A_n be a closed set contained in $(n, n+1)$. Suppose that $f: \mathbf{R} \rightarrow \mathbf{R}$ satisfies $\int_{A_n} f = (-1)^n/n$ and $f = 0$ for $x \notin \text{any } A_n$. Find two partitions of unity Φ and Ψ such that $\sum_{\varphi \in \Phi} \int_{\mathbf{R}} \varphi \cdot f$ and $\sum_{\psi \in \Psi} \int_{\mathbf{R}} \psi \cdot f$ converge absolutely to different values.

CHANGE OF VARIABLE

If $g: [a,b] \rightarrow \mathbf{R}$ is continuously differentiable and $f: \mathbf{R} \rightarrow \mathbf{R}$ is continuous, then, as is well known,

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g) \cdot g'.$$

The proof is very simple: if $F' = f$, then $(F \circ g)' = (f \circ g) \cdot g'$; thus the left side is $F(g(b)) - F(g(a))$, while the right side is $F \circ g(b) - F \circ g(a) = F(g(b)) - F(g(a))$.

We leave it to the reader to show that if g is 1-1, then the above formula can be written

$$\int_{g((a,b))} f = \int_{(a,b)} f \circ g \cdot |g'|.$$

(Consider separately the cases where g is increasing and where g is decreasing.) The generalization of this formula to higher dimensions is by no means so trivial.

3-13 Theorem. *Let $A \subset \mathbf{R}^n$ be an open set and $g: A \rightarrow \mathbf{R}^n$ a 1-1, continuously differentiable function such that $\det g'(x) \neq 0$ for all $x \in A$. If $f: g(A) \rightarrow \mathbf{R}$ is integrable, then*

$$\int_{g(A)} f = \int_A (f \circ g) |\det g'|.$$

Proof. We begin with some important reductions.

1. Suppose there is an admissible cover \mathcal{O} for A such that for each $U \in \mathcal{O}$ and any integrable f we have

$$\int_{g(U)} f = \int_U (f \circ g) |\det g'|.$$

Then the theorem is true for all of A . (Since g is automatically 1-1 in an open set around each point, it is not surprising that this is the only part of the proof using the fact that g is 1-1 on all of A .)

Proof of (1). The collection of all $g(U)$ is an open cover of $g(A)$. Let Φ be a partition of unity subordinate to this cover. If $\varphi = 0$ outside of $g(U)$, then, since g is 1-1, we have $(\varphi \cdot f) \circ g$

= 0 outside of U . Therefore the equation

$$\int_{g(U)} \varphi \cdot f = \int_U [(\varphi \cdot f) \circ g] |\det g'|.$$

can be written

$$\int_{g(A)} \varphi \cdot f = \int_A [(\varphi \cdot f) \circ g] |\det g'|.$$

Hence

$$\begin{aligned} \int_{g(A)} f &= \sum_{\varphi \in \Phi} \int_{g(A)} \varphi \cdot f = \sum_{\varphi \in \Phi} \int_A [(\varphi \cdot f) \circ g] |\det g'| \\ &= \sum_{\varphi \in \Phi} \int_A (\varphi \circ g)(f \circ g) |\det g'| \\ &= \int_A (f \circ g) |\det g'|. \end{aligned}$$

Remark. The theorem also follows from the assumption that

$$\int_V f = \int_{g^{-1}(V)} (f \circ g) |\det g'|$$

for V in some admissible cover of $g(A)$. This follows from (1) applied to g^{-1} .

2. It suffices to prove the theorem for the function $f = 1$.

Proof of (2). If the theorem holds for $f = 1$, it holds for constant functions. Let V be a rectangle in $g(A)$ and P a partition of V . For each subrectangle S of P let f_S be the constant function $m_S(f)$. Then

$$\begin{aligned} L(f, P) &= \sum_S m_S(f) \cdot v(S) = \sum_S \int_{\text{int } S} f_S \\ &= \sum_S \int_{g^{-1}(\text{int } S)} (f_S \circ g) |\det g'| \leq \sum_S \int_{g^{-1}(\text{int } S)} (f \circ g) |\det g'| \\ &\leq \int_{g^{-1}(V)} (f \circ g) |\det g'|. \end{aligned}$$

Since $\int_V f$ is the least upper bound of all $L(f, P)$, this proves that $\int_V f \leq \int_{g^{-1}(V)} (f \circ g) |\det g'|$. A similar argument, letting $f_S = M_S(f)$, shows that $\int_V f \geq \int_{g^{-1}(V)} (f \circ g) |\det g'|$. The result now follows from the above Remark.

3. If the theorem is true for $g: A \rightarrow \mathbf{R}^n$ and for $h: B \rightarrow \mathbf{R}^n$, where $g(A) \subset B$, then it is true for $h \circ g: A \rightarrow \mathbf{R}^n$.

Proof of (3).

$$\begin{aligned} \int_{h \circ g(A)} f &= \int_{h(g(A))} f = \int_{g(A)} (f \circ h) |\det h'| \\ &= \int_A [(f \circ h) \circ g] \cdot [|\det h'| \circ g] \cdot |\det g'| \\ &= \int_A f \circ (h \circ g) |\det (h \circ g)'|. \end{aligned}$$

4. The theorem is true if g is a linear transformation.

Proof of (4). By (1) and (2) it suffices to show for any open rectangle U that

$$\int_{g(U)} 1 = \int_U |\det g'|.$$

This is Problem 3-35.

Observations (3) and (4) together show that we may assume for any particular $a \in A$ that $g'(a)$ is the identity matrix: in fact, if T is the linear transformation $Dg(a)$, then $(T^{-1} \circ g)'(a) = I$; since the theorem is true for T , if it is true for $T^{-1} \circ g$ it will be true for g .

We are now prepared to give the proof, which preceeds by induction on n . The remarks before the statement of the theorem, together with (1) and (2), prove the case $n = 1$. Assuming the theorem in dimension $n - 1$, we prove it in dimension n . For each $a \in A$ we need only find an open set U with $a \in U \subset A$ for which the theorem is true. Moreover we may assume that $g'(a) = I$.

Define $h: A \rightarrow \mathbf{R}^n$ by $h(x) = (g^1(x), \dots, g^{n-1}(x), x^n)$. Then $h'(a) = I$. Hence in some open U' with $a \in U' \subset A$, the function h is 1-1 and $\det h'(x) \neq 0$. We can thus define $k: h(U') \rightarrow \mathbf{R}^n$ by $k(x) = (x^1, \dots, x^{n-1}, g^n(h^{-1}(x)))$ and $g = k \circ h$. We have thus expressed g as the composition

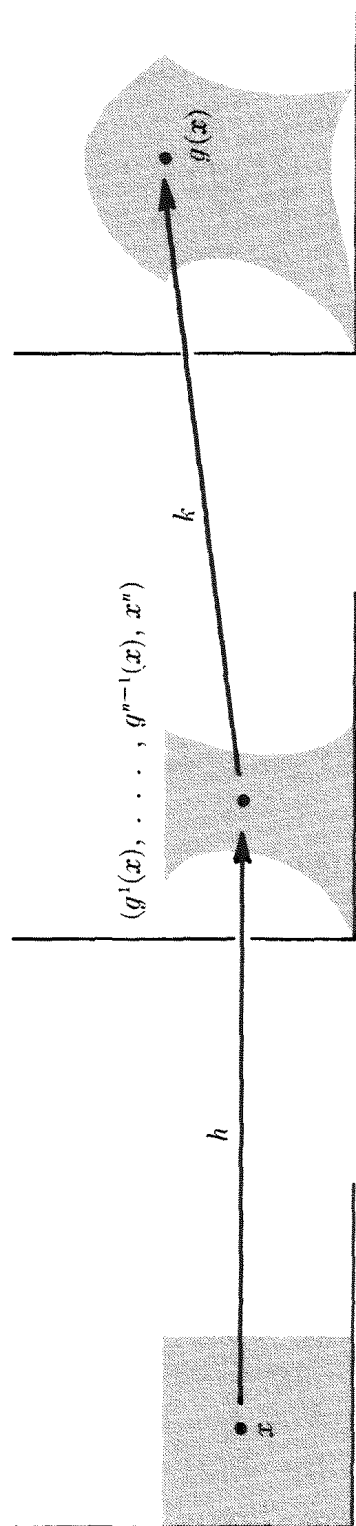


FIGURE 3-3

of two maps, each of which changes fewer than n coordinates (Figure 3-3).

We must attend to a few details to ensure that k is a function of the proper sort. Since

$$(g^n \circ h^{-1})'(h(a)) = (g^n)'(a) \cdot [h'(a)]^{-1} = (g^n)'(a),$$

we have $D_n(g^n \circ h^{-1})(h(a)) = D_n g^n(a) = 1$, so that $k'(h(a)) = I$. Thus in some open set V with $h(a) \in V \subset h(U')$, the function k is 1-1 and $\det k'(x) \neq 0$. Letting $U = k^{-1}(V)$ we now have $g = k \circ h$, where $h: U \rightarrow \mathbf{R}^n$ and $k: V \rightarrow \mathbf{R}^n$ and $h(U) \subset V$. By (3) it suffices to prove the theorem for h and k . We give the proof for h ; the proof for k is similar and easier.

Let $W \subset U$ be a rectangle of the form $D \times [a_n, b_n]$, where D is a rectangle in \mathbf{R}^{n-1} . By Fubini's theorem

$$\int_{h(W)} 1 = \int_{[a_n, b_n]} \left(\int_{h(D \times \{x^n\})} 1 \, dx^1 \cdots dx^{n-1} \right) dx^n.$$

Let $h_{x^n}: D \rightarrow \mathbf{R}^{n-1}$ be defined by $h_{x^n}(x^1, \dots, x^{n-1}) = (g^1(x^1, \dots, x^n), \dots, g^{n-1}(x^1, \dots, x^n))$. Then each h_{x^n} is clearly 1-1 and

$$\det (h_{x^n})'(x^1, \dots, x^{n-1}) = \det h'(x^1, \dots, x^n) \neq 0.$$

Moreover

$$\int_{h(D \times \{x^n\})} 1 \, dx^1 \cdots dx^{n-1} = \int_{h_{x^n}(D)} 1 \, dx^1 \cdots dx^{n-1}.$$

Applying the theorem in the case $n - 1$ therefore gives

$$\begin{aligned} \int_{h(W)} 1 &= \int_{[a_n, b_n]} \left(\int_{h_{x^n}(D)} 1 \, dx^1 \cdots dx^{n-1} \right) dx^n \\ &= \int_{[a_n, b_n]} \left(\int_D |\det (h_{x^n})'(x^1, \dots, x^{n-1})| \, dx^1 \cdots dx^{n-1} \right) dx^n \\ &= \int_{[a_n, b_n]} \left(\int_D |\det h'(x^1, \dots, x^n)| \, dx^1 \cdots dx^{n-1} \right) dx^n \\ &= \int_W |\det h'|. \quad \blacksquare \end{aligned}$$

The condition $\det g'(x) \neq 0$ may be eliminated from the

hypotheses of Theorem 3-13 by using the following theorem, which often plays an unexpected role.

3-14. Theorem (Sard's Theorem). *Let $g: A \rightarrow \mathbf{R}^n$ be continuously differentiable, where $A \subset \mathbf{R}^n$ is open, and let $B = \{x \in A: \det g'(x) = 0\}$. Then $g(B)$ has measure 0.*

Proof. Let $U \subset A$ be a closed rectangle such that all sides of U have length l , say. Let $\varepsilon > 0$. If N is sufficiently large and U is divided into N^n rectangles, with sides of length l/N , then for each of these rectangles S , if $x \in S$ we have

$$|Dg(x)(y - x) - g(y) + g(x)| < \varepsilon |x - y| \leq \varepsilon \sqrt{n} (l/N)$$

for all $y \in S$. If S intersects B we can choose $x \in S \cap B$; since $\det g'(x) = 0$, the set $\{Dg(x)(y - x): y \in S\}$ lies in an $(n - 1)$ -dimensional subspace V of \mathbf{R}^n . Therefore the set $\{g(y) - g(x): y \in S\}$ lies within $\varepsilon \sqrt{n} (l/N)$ of V , so that $\{g(y): y \in S\}$ lies within $\varepsilon \sqrt{n} (l/N)$ of the $(n - 1)$ -plane $V + g(x)$. On the other hand, by Lemma 2-10 there is a number M such that

$$|g(x) - g(y)| < M |x - y| \leq M \sqrt{n} (l/N).$$

Thus, if S intersects B , the set $\{g(y): y \in S\}$ is contained in a cylinder whose height is $< 2\varepsilon \sqrt{n} (l/N)$ and whose base is an $(n - 1)$ -dimensional sphere of radius $< M \sqrt{n} (l/N)$. This cylinder has volume $< C(l/N)^n \varepsilon$ for some constant C . There are at most N^n such rectangles S , so $g(U \cap B)$ lies in a set of volume $< C(l/N)^n \cdot \varepsilon \cdot N^n = Cl^n \cdot \varepsilon$. Since this is true for all $\varepsilon > 0$, the set $g(U \cap B)$ has measure 0. Since (Problem 3-13) we can cover all of A with a sequence of such rectangles U , the desired result follows from Theorem 3-4. ■

Theorem 3-14 is actually only the easy part of Sard's Theorem. The statement and proof of the deeper result will be found in [17], page 47.

Problems. 3-39. Use Theorem 3-14 to prove Theorem 3-13 without the assumption $\det g'(x) \neq 0$.

3-40. If $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $\det g'(x) \neq 0$, prove that in some open set containing x we can write $g = T \circ g_n \circ \cdots \circ g_1$, where g_i is of the form $g_i(x) = (x^1, \dots, f_i(x), \dots, x^n)$, and T is a linear transformation. Show that we can write $g = g_n \circ \cdots \circ g_1$ if and only if $g'(x)$ is a diagonal matrix.

3-41. Define $f: \{r: r > 0\} \times (0, 2\pi) \rightarrow \mathbf{R}^2$ by $f(r, \theta) = (r \cos \theta, r \sin \theta)$.

(a) Show that f is 1-1, compute $f'(r, \theta)$, and show that $\det f'(r, \theta) \neq 0$ for all (r, θ) . Show that $f(\{r: r > 0\} \times (0, 2\pi))$ is the set A of Problem 2-23.

(b) If $P = f^{-1}$, show that $P(x, y) = (r(x, y), \theta(x, y))$, where

$$r(x, y) = \sqrt{x^2 + y^2},$$

$$\theta(x, y) = \begin{cases} \arctan y/x & x > 0, y > 0, \\ \pi + \arctan y/x & x < 0, \\ 2\pi + \arctan y/x & x > 0, y < 0, \\ \pi/2 & x = 0, y > 0, \\ 3\pi/2 & x = 0, y < 0. \end{cases}$$

(Here \arctan denotes the inverse of the function $\tan: (-\pi/2, \pi/2) \rightarrow \mathbf{R}$.) Find $P'(x, y)$. The function P is called the **polar coordinate system** on A .

(c) Let $C \subset A$ be the region between the circles of radii r_1 and r_2 and the half-lines through 0 which make angles of θ_1 and θ_2 with the x -axis. If $h: C \rightarrow \mathbf{R}$ is integrable and $h(x, y) = g(r(x, y), \theta(x, y))$, show that

$$\int_C h = \int_{r_1}^{r_2} \int_{\theta_1}^{\theta_2} r g(r, \theta) d\theta dr.$$

If $B_r = \{(x, y): x^2 + y^2 \leq r^2\}$, show that

$$\int_{B_r} h = \int_0^r \int_0^{2\pi} r g(r, \theta) d\theta dr.$$

(d) If $C_r = [-r, r] \times [-r, r]$, show that

$$\int_{B_r} e^{-(x^2+y^2)} dx dy = \pi(1 - e^{-r^2})$$

and

$$\int_{C_r} e^{-(x^2+y^2)} dx dy = \left(\int_{-r}^r e^{-x^2} dx \right)^2.$$

(e) Prove that

$$\lim_{r \rightarrow \infty} \int_{B_r} e^{-(x^2+y^2)} dx dy = \lim_{r \rightarrow \infty} \int_{C_r} e^{-(x^2+y^2)} dx dy$$

and conclude that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

“A mathematician is one to whom *that* is as obvious as that twice two makes four is to you. Liouville was a mathematician.”

—LORD KELVIN

4

Integration on Chains

ALGEBRAIC PRELIMINARIES

If V is a vector space (over \mathbf{R}), we will denote the k -fold product $V \times \cdots \times V$ by V^k . A function $T: V^k \rightarrow \mathbf{R}$ is called **multilinear** if for each i with $1 \leq i \leq k$ we have

$$\begin{aligned} T(v_1, \dots, v_i + v_i', \dots, v_k) &= T(v_1, \dots, v_i, \dots, v_k) \\ &\quad + T(v_1, \dots, v_i', \dots, v_k), \\ T(v_1, \dots, av_i, \dots, v_k) &= aT(v_1, \dots, v_i, \dots, v_k). \end{aligned}$$

A multilinear function $T: V^k \rightarrow \mathbf{R}$ is called a **k -tensor** on V and the set of all k -tensors, denoted $\mathfrak{J}^k(V)$, becomes a vector space (over \mathbf{R}) if for $S, T \in \mathfrak{J}^k(V)$ and $a \in \mathbf{R}$ we define

$$\begin{aligned} (S + T)(v_1, \dots, v_k) &= S(v_1, \dots, v_k) + T(v_1, \dots, v_k), \\ (aS)(v_1, \dots, v_k) &= a \cdot S(v_1, \dots, v_k). \end{aligned}$$

There is also an operation connecting the various spaces $\mathfrak{J}^k(V)$. If $S \in \mathfrak{J}^k(V)$ and $T \in \mathfrak{J}^l(V)$, we define the **tensor product** $S \otimes T \in \mathfrak{J}^{k+l}(V)$ by

$$\begin{aligned} S \otimes T(v_1, \dots, v_k, v_{k+1}, \dots, v_{k+l}) \\ = S(v_1, \dots, v_k) \cdot T(v_{k+1}, \dots, v_{k+l}). \end{aligned}$$

Note that the order of the factors S and T is crucial here since $S \otimes T$ and $T \otimes S$ are far from equal. The following properties of \otimes are left as easy exercises for the reader.

$$\begin{aligned}(S_1 + S_2) \otimes T &= S_1 \otimes T + S_2 \otimes T, \\ S \otimes (T_1 + T_2) &= S \otimes T_1 + S \otimes T_2, \\ (aS) \otimes T &= S \otimes (aT) = a(S \otimes T), \\ (S \otimes T) \otimes U &= S \otimes (T \otimes U).\end{aligned}$$

Both $(S \otimes T) \otimes U$ and $S \otimes (T \otimes U)$ are usually denoted simply $S \otimes T \otimes U$; higher-order products $T_1 \otimes \cdots \otimes T_r$ are defined similarly.

The reader has probably already noticed that $\mathfrak{I}^1(V)$ is just the dual space V^* . The operation \otimes allows us to express the other vector spaces $\mathfrak{I}^k(V)$ in terms of $\mathfrak{I}^1(V)$.

4-1 Theorem. *Let v_1, \dots, v_n be a basis for V , and let $\varphi_1, \dots, \varphi_n$ be the dual basis, $\varphi_i(v_j) = \delta_{ij}$. Then the set of all k -fold tensor products*

$$\varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k} \quad 1 \leq i_1, \dots, i_k \leq n$$

is a basis for $\mathfrak{I}^k(V)$, which therefore has dimension n^k .

Proof. Note that

$$\begin{aligned}\varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k}(v_{j_1}, \dots, v_{j_k}) &= \delta_{i_1, j_1} \cdots \delta_{i_k, j_k} \\ &= \begin{cases} 1 & \text{if } j_1 = i_1, \dots, j_k = i_k, \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

If w_1, \dots, w_k are k vectors with $w_i = \sum_{j=1}^n a_{ij} v_j$ and T is in $\mathfrak{I}^k(V)$, then

$$\begin{aligned}T(w_1, \dots, w_k) &= \sum_{j_1, \dots, j_k=1}^n a_{1, j_1} \cdots a_{k, j_k} T(v_{j_1}, \dots, v_{j_k}) \\ &= \sum_{i_1, \dots, i_k=1}^n T(v_{i_1}, \dots, v_{i_k}) \cdot \varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k}(w_1, \dots, w_k).\end{aligned}$$

Thus $T = \sum_{i_1, \dots, i_k=1}^n T(v_{i_1}, \dots, v_{i_k}) \cdot \varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k}$.

Consequently the $\varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k}$ span $\mathfrak{I}^k(V)$.

Suppose now that there are numbers a_{i_1, \dots, i_k} such that

$$\sum_{i_1, \dots, i_k=1}^n a_{i_1, \dots, i_k} \cdot \varphi_{i_1} \otimes \dots \otimes \varphi_{i_k} = 0.$$

Applying both sides of this equation to $(v_{j_1}, \dots, v_{j_k})$ yields $a_{j_1, \dots, j_k} = 0$. Thus the $\varphi_{i_1} \otimes \dots \otimes \varphi_{i_k}$ are linearly independent. ■

One important construction, familiar for the case of dual spaces, can also be made for tensors. If $f: V \rightarrow W$ is a linear transformation, a linear transformation $f^*: \mathfrak{J}^k(W) \rightarrow \mathfrak{J}^k(V)$ is defined by

$$f^*T(v_1, \dots, v_k) = T(f(v_1), \dots, f(v_k))$$

for $T \in \mathfrak{J}^k(W)$ and $v_1, \dots, v_k \in V$. It is easy to verify that $f^*(S \otimes T) = f^*S \otimes f^*T$.

The reader is already familiar with certain tensors, aside from members of V^* . The first example is the inner product $\langle, \rangle \in \mathfrak{J}^2(\mathbf{R}^n)$. On the grounds that any good mathematical commodity is worth generalizing, we define an **inner product** on V to be a 2-tensor T such that T is **symmetric**, that is $T(v, w) = T(w, v)$ for $v, w \in V$ and such that T is **positive-definite**, that is, $T(v, v) > 0$ if $v \neq 0$. We distinguish \langle, \rangle as the **usual inner product** on \mathbf{R}^n . The following theorem shows that our generalization is not too general.

4-2 Theorem. *If T is an inner product on V , there is a basis v_1, \dots, v_n for V such that $T(v_i, v_j) = \delta_{ij}$. (Such a basis is called **orthonormal** with respect to T .) Consequently there is an isomorphism $f: \mathbf{R}^n \rightarrow V$ such that $T(f(x), f(y)) = \langle x, y \rangle$ for $x, y \in \mathbf{R}^n$. In other words $f^*T = \langle, \rangle$.*

Proof. Let w_1, \dots, w_n be any basis for V . Define

$$\begin{aligned} w_1' &= w_1, \\ w_2' &= w_2 - \frac{T(w_1', w_2)}{T(w_1', w_1')} \cdot w_1', \\ w_3' &= w_3 - \frac{T(w_1', w_3)}{T(w_1', w_1')} \cdot w_1' - \frac{T(w_2', w_3)}{T(w_2', w_2')} \cdot w_2', \\ &\text{etc.} \end{aligned}$$

It is easy to check that $T(w_i', w_j') = 0$ if $i \neq j$ and $w_i' \neq 0$ so that $T(w_i', w_i') > 0$. Now define $v_i = w_i' / \sqrt{T(w_i', w_i')}$. The isomorphism f may be defined by $f(e_i) = v_i$. ■

Despite its importance, the inner product plays a far lesser role than another familiar, seemingly ubiquitous function, the tensor $\det \in \mathfrak{I}^n(\mathbb{R}^n)$. In attempting to generalize this function, we recall that interchanging two rows of a matrix changes the sign of its determinant. This suggests the following definition. A k -tensor $\omega \in \mathfrak{I}^k(V)$ is called **alternating** if

$$\begin{aligned} \omega(v_1, \dots, v_i, \dots, v_j, \dots, v_k) \\ = -\omega(v_1, \dots, v_j, \dots, v_i, \dots, v_k) \end{aligned} \quad \text{for all } v_1, \dots, v_k \in V.$$

(In this equation v_i and v_j are interchanged and all other v 's are left fixed.) The set of all alternating k -tensors is clearly a subspace $\Lambda^k(V)$ of $\mathfrak{I}^k(V)$. Since it requires considerable work to produce the determinant, it is not surprising that alternating k -tensors are difficult to write down. There is, however, a uniform way of expressing all of them. Recall that the sign of a permutation σ , denoted $\text{sgn } \sigma$, is $+1$ if σ is even and -1 if σ is odd. If $T \in \mathfrak{I}^k(V)$, we define $\text{Alt}(T)$ by

$$\text{Alt}(T)(v_1, \dots, v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot T(v_{\sigma(1)}, \dots, v_{\sigma(k)}),$$

where S_k is the set of all permutations of the numbers 1 to k .

4-3 Theorem

- (1) If $T \in \mathfrak{I}^k(V)$, then $\text{Alt}(T) \in \Lambda^k(V)$.
- (2) If $\omega \in \Lambda^k(V)$, then $\text{Alt}(\omega) = \omega$.
- (3) If $T \in \mathfrak{I}^k(V)$, then $\text{Alt}(\text{Alt}(T)) = \text{Alt}(T)$.

Proof

- (1) Let (i, j) be the permutation that interchanges i and j and leaves all other numbers fixed. If $\sigma \in S_k$, let $\sigma' = \sigma \cdot (i, j)$. Then

$$\begin{aligned}
& \text{Alt}(T)(v_1, \dots, v_j, \dots, v_i, \dots, v_k) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot T(v_{\sigma(1)}, \dots, v_{\sigma(j)}, \dots, v_{\sigma(i)}, \dots, v_{\sigma(k)}) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot T(v_{\sigma'(1)}, \dots, v_{\sigma'(i)}, \dots, v_{\sigma'(j)}, \dots, v_{\sigma'(k)}) \\
&= \frac{1}{k!} \sum_{\sigma' \in S_k} -\text{sgn } \sigma' \cdot T(v_{\sigma'(1)}, \dots, v_{\sigma'(k)}) \\
&= -\text{Alt}(T)(v_1, \dots, v_k).
\end{aligned}$$

(2) If $\omega \in \Lambda^k(V)$, and $\sigma = (i, j)$, then $\omega(v_{\sigma(1)}, \dots, v_{\sigma(k)}) = \text{sgn } \sigma \cdot \omega(v_1, \dots, v_k)$. Since every σ is a product of permutations of the form (i, j) , this equation holds of all σ . Therefore

$$\begin{aligned}
\text{Alt}(\omega)(v_1, \dots, v_k) &= \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot \omega(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn } \sigma \cdot \text{sgn } \sigma \cdot \omega(v_1, \dots, v_k) \\
&= \omega(v_1, \dots, v_k).
\end{aligned}$$

(3) follows immediately from (1) and (2). ■

To determine the dimensions of $\Lambda^k(V)$, we would like a theorem analogous to Theorem 4-1. Of course, if $\omega \in \Lambda^k(V)$ and $\eta \in \Lambda^l(V)$, then $\omega \otimes \eta$ is usually not in $\Lambda^{k+l}(V)$. We will therefore define a new product, the **wedge product** $\omega \wedge \eta \in \Lambda^{k+l}(V)$ by

$$\omega \wedge \eta = \frac{(k+l)!}{k! l!} \text{Alt}(\omega \otimes \eta).$$

(The reason for the strange coefficient will appear later.) The following properties of \wedge are left as an exercise for the reader:

$$\begin{aligned}
(\omega_1 + \omega_2) \wedge \eta &= \omega_1 \wedge \eta + \omega_2 \wedge \eta, \\
\omega \wedge (\eta_1 + \eta_2) &= \omega \wedge \eta_1 + \omega \wedge \eta_2, \\
a\omega \wedge \eta &= \omega \wedge a\eta = a(\omega \wedge \eta), \\
\omega \wedge \eta &= (-1)^{kl} \eta \wedge \omega, \\
f^*(\omega \wedge \eta) &= f^*(\omega) \wedge f^*(\eta).
\end{aligned}$$

The equation $(\omega \wedge \eta) \wedge \theta = \omega \wedge (\eta \wedge \theta)$ is true but requires more work.

4-4 Theorem

(1) If $S \in \mathfrak{I}^k(V)$ and $T \in \mathfrak{I}^l(V)$ and $\text{Alt}(S) = 0$, then

$$\text{Alt}(S \otimes T) = \text{Alt}(T \otimes S) = 0.$$

(2) $\text{Alt}(\text{Alt}(\omega \otimes \eta) \otimes \theta) = \text{Alt}(\omega \otimes \eta \otimes \theta)$
 $= \text{Alt}(\omega \otimes \text{Alt}(\eta \otimes \theta)).$

(3) If $\omega \in \Lambda^k(V)$, $\eta \in \Lambda^l(V)$, and $\theta \in \Lambda^m(V)$, then

$$\begin{aligned} (\omega \wedge \eta) \wedge \theta &= \omega \wedge (\eta \wedge \theta) \\ &= \frac{(k + l + m)!}{k! l! m!} \text{Alt}(\omega \otimes \eta \otimes \theta). \end{aligned}$$

Proof

(1)

$$\begin{aligned} &(k + l)! \text{Alt}(S \otimes T)(v_1, \dots, v_{k+l}) \\ &= \sum_{\sigma \in S_{k+l}} \text{sgn } \sigma \cdot S(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot T(v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}). \end{aligned}$$

If $G \subset S_{k+l}$ consists of all σ which leave $k + 1, \dots, k + l$ fixed, then

$$\begin{aligned} &\sum_{\sigma \in G} \text{sgn } \sigma \cdot S(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot T(v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}) \\ &= \left[\sum_{\sigma' \in S_k} \text{sgn } \sigma' \cdot S(v_{\sigma'(1)}, \dots, v_{\sigma'(k)}) \right] \cdot T(v_{k+1}, \dots, v_{k+l}) \\ &= 0. \end{aligned}$$

Suppose now that $\sigma_0 \notin G$. Let $G \cdot \sigma_0 = \{\sigma \cdot \sigma_0 : \sigma \in G\}$ and let $v_{\sigma_0(1)}, \dots, v_{\sigma_0(k+l)} = w_1, \dots, w_{k+l}$. Then

$$\begin{aligned} &\sum_{\sigma \in G \cdot \sigma_0} \text{sgn } \sigma \cdot S(v_{\sigma(1)}, \dots, v_{\sigma(k)}) \cdot T(v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}) \\ &= \left[\text{sgn } \sigma_0 \cdot \sum_{\sigma' \in G} \text{sgn } \sigma' \cdot S(w_{\sigma'(1)}, \dots, w_{\sigma'(k)}) \right] \\ &\quad \cdot T(w_{k+1}, \dots, w_{k+l}) \\ &= 0. \end{aligned}$$

Notice that $G \cap G \cdot \sigma_0 = \emptyset$. In fact, if $\sigma \in G \cap G \cdot \sigma_0$, then $\sigma = \sigma' \cdot \sigma_0$ for some $\sigma' \in G$ and $\sigma_0 = \sigma \cdot (\sigma')^{-1} \in G$, a contradiction. We can then continue in this way, breaking S_{k+l} up into disjoint subsets; the sum over each subset is 0, so that the sum over S_{k+l} is 0. The relation $\text{Alt}(T \otimes S) = 0$ is proved similarly.

(2) We have

$$\text{Alt}(\text{Alt}(\eta \otimes \theta) - \eta \otimes \theta) = \text{Alt}(\eta \otimes \theta) - \text{Alt}(\eta \otimes \theta) = 0.$$

Hence by (1) we have

$$\begin{aligned} 0 &= \text{Alt}(\omega \otimes [\text{Alt}(\eta \otimes \theta) - \eta \otimes \theta]) \\ &= \text{Alt}(\omega \otimes \text{Alt}(\eta \otimes \theta)) - \text{Alt}(\omega \otimes \eta \otimes \theta). \end{aligned}$$

The other equality is proved similarly.

$$\begin{aligned} (3) \quad (\omega \wedge \eta) \wedge \theta &= \frac{(k+l+m)!}{(k+l)!m!} \text{Alt}((\omega \wedge \eta) \otimes \theta) \\ &= \frac{(k+l+m)!}{(k+l)!m!} \frac{(k+l)!}{k!l!} \text{Alt}(\omega \otimes \eta \otimes \theta). \end{aligned}$$

The other equality is proved similarly. ■

Naturally $\omega \wedge (\eta \wedge \theta)$ and $(\omega \wedge \eta) \wedge \theta$ are both denoted simply $\omega \wedge \eta \wedge \theta$, and higher-order products $\omega_1 \wedge \cdots \wedge \omega_r$ are defined similarly. If v_1, \dots, v_n is a basis for V and $\varphi_1, \dots, \varphi_n$ is the dual basis, a basis for $\Lambda^k(V)$ can now be constructed quite easily.

4-5 Theorem. *The set of all*

$$\varphi_{i_1} \wedge \cdots \wedge \varphi_{i_k} \quad 1 \leq i_1 < i_2 < \cdots < i_k \leq n$$

is a basis for $\Lambda^k(V)$, which therefore has dimension

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Proof. If $\omega \in \Lambda^k(V) \subset \mathcal{J}^k(V)$, then we can write

$$\omega = \sum_{i_1, \dots, i_k} a_{i_1, \dots, i_k} \varphi_{i_1} \otimes \cdots \otimes \varphi_{i_k}.$$

Thus

$$\omega = \text{Alt}(\omega) = \sum_{i_1, \dots, i_k} a_{i_1, \dots, i_k} \text{Alt}(\varphi_{i_1} \otimes \dots \otimes \varphi_{i_k}).$$

Since each $\text{Alt}(\varphi_{i_1} \otimes \dots \otimes \varphi_{i_k})$ is a constant times one of the $\varphi_{i_1} \wedge \dots \wedge \varphi_{i_k}$, these elements span $\Lambda^k(V)$. Linear independence is proved as in Theorem 4-1 (cf. Problem 4-1). ■

If V has dimension n , it follows from Theorem 4-5 that $\Lambda^n(V)$ has dimension 1. Thus all alternating n -tensors on V are multiples of any non-zero one. Since the determinant is an example of such a member of $\Lambda^n(\mathbf{R}^n)$, it is not surprising to find it in the following theorem.

4-6 Theorem. *Let v_1, \dots, v_n be a basis for V , and let $\omega \in \Lambda^n(V)$. If $w_i = \sum_{j=1}^n a_{ij}v_j$ are n vectors in V , then*

$$\omega(w_1, \dots, w_n) = \det(a_{ij}) \cdot \omega(v_1, \dots, v_n).$$

Proof. Define $\eta \in \mathcal{T}^n(\mathbf{R}^n)$ by

$$\begin{aligned} \eta((a_{11}, \dots, a_{1n}), \dots, (a_{n1}, \dots, a_{nn})) \\ = \omega(\sum a_{1j}v_j, \dots, \sum a_{nj}v_j). \end{aligned}$$

Clearly $\eta \in \Lambda^n(\mathbf{R}^n)$ so $\eta = \lambda \cdot \det$ for some $\lambda \in \mathbf{R}$ and $\lambda = \eta(e_1, \dots, e_n) = \omega(v_1, \dots, v_n)$. ■

Theorem 4-6 shows that a non-zero $\omega \in \Lambda^n(V)$ splits the bases of V into two disjoint groups, those with $\omega(v_1, \dots, v_n) > 0$ and those for which $\omega(v_1, \dots, v_n) < 0$; if v_1, \dots, v_n and w_1, \dots, w_n are two bases and $A = (a_{ij})$ is defined by $w_i = \sum a_{ij}v_j$, then v_1, \dots, v_n and w_1, \dots, w_n are in the same group if and only if $\det A > 0$. This criterion is independent of ω and can always be used to divide the bases of V into two disjoint groups. Either of these two groups is called an **orientation** for V . The orientation to which a basis v_1, \dots, v_n belongs is denoted $[v_1, \dots, v_n]$ and the

other orientation is denoted $-[v_1, \dots, v_n]$. In \mathbf{R}^n we define the **usual orientation** as $[e_1, \dots, e_n]$.

The fact that $\dim \Lambda^n(\mathbf{R}^n) = 1$ is probably not new to you, since \det is often defined as the unique element $\omega \in \Lambda^n(\mathbf{R}^n)$ such that $\omega(e_1, \dots, e_n) = 1$. For a general vector space V there is no extra criterion of this sort to distinguish a particular $\omega \in \Lambda^n(V)$. Suppose, however, that an inner product T for V is given. If v_1, \dots, v_n and w_1, \dots, w_n are two bases which are orthonormal with respect to T , and the matrix $A = (a_{ij})$ is defined by $w_i = \sum_{j=1}^n a_{ij}v_j$, then

$$\begin{aligned} \delta_{ij} = T(w_i, w_j) &= \sum_{k,l=1}^n a_{ik}a_{jl}T(v_k, v_l) \\ &= \sum_{k=1}^n a_{ik}a_{jk}. \end{aligned}$$

In other words, if A^T denotes the transpose of the matrix A , then we have $A \cdot A^T = I$, so $\det A = \pm 1$. It follows from Theorem 4-6 that if $\omega \in \Lambda^n(V)$ satisfies $\omega(v_1, \dots, v_n) = \pm 1$, then $\omega(w_1, \dots, w_n) = \pm 1$. If an orientation μ for V has also been given, it follows that there is a unique $\omega \in \Lambda^n(V)$ such that $\omega(v_1, \dots, v_n) = 1$ whenever v_1, \dots, v_n is an orthonormal basis such that $[v_1, \dots, v_n] = \mu$. This unique ω is called the **volume element** of V , determined by the inner product T and orientation μ . Note that \det is the volume element of \mathbf{R}^n determined by the usual inner product and usual orientation, and that $|\det(v_1, \dots, v_n)|$ is the volume of the parallelipiped spanned by the line segments from 0 to each of v_1, \dots, v_n .

We conclude this section with a construction which we will restrict to \mathbf{R}^n . If $v_1, \dots, v_{n-1} \in \mathbf{R}^n$ and φ is defined by

$$\varphi(w) = \det \begin{pmatrix} v_1 \\ \vdots \\ v_{n-1} \\ w \end{pmatrix},$$

then $\varphi \in \Lambda^1(\mathbf{R}^n)$; therefore there is a unique $z \in \mathbf{R}^n$ such that

$$\langle w, z \rangle = \varphi(w) = \det \begin{pmatrix} v_1 \\ \vdots \\ v_{n-1} \\ w \end{pmatrix}$$

This z is denoted $v_1 \times \cdots \times v_{n-1}$ and called the **cross product** of v_1, \dots, v_{n-1} . The following properties are immediate from the definition:

$$\begin{aligned} v_{\sigma(1)} \times \cdots \times v_{\sigma(n-1)} &= \operatorname{sgn} \sigma \cdot v_1 \times \cdots \times v_{n-1}, \\ v_1 \times \cdots \times av_i \times \cdots \times v_{n-1} &= a \cdot (v_1 \times \cdots \times v_{n-1}), \\ v_1 \times \cdots \times (v_i + v_i') \times \cdots \times v_{n-1} \\ &= v_1 \times \cdots \times v_i \times \cdots \times v_{n-1} \\ &\quad + v_1 \times \cdots \times v_i' \times \cdots \times v_{n-1}. \end{aligned}$$

It is uncommon in mathematics to have a “product” that depends on more than two factors. In the case of two vectors $v, w \in \mathbf{R}^3$, we obtain a more conventional looking product, $v \times w \in \mathbf{R}^3$. For this reason it is sometimes maintained that the cross product can be defined only in \mathbf{R}^3 .

Problems. 4-1.* Let e_1, \dots, e_n be the usual basis of \mathbf{R}^n and let $\varphi_1, \dots, \varphi_n$ be the dual basis.

(a) Show that $\varphi_{i_1} \wedge \cdots \wedge \varphi_{i_k} (e_{i_1}, \dots, e_{i_k}) = 1$. What would the right side be if the factor $(k+l)!/k!l!$ did not appear in the definition of \wedge ?

(b) Show that $\varphi_{i_1} \wedge \cdots \wedge \varphi_{i_k} (v_1, \dots, v_k)$ is the determinant

of the $k \times k$ minor of $\begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}$ obtained by selecting columns

i_1, \dots, i_k .

4-2. If $f: V \rightarrow V$ is a linear transformation and $\dim V = n$, then $f^*: \Lambda^n(V) \rightarrow \Lambda^n(V)$ must be multiplication by some constant c . Show that $c = \det f$.

- 4-3. If $\omega \in \Lambda^n(V)$ is the volume element determined by T and μ , and $w_1, \dots, w_n \in V$, show that

$$|\omega(w_1, \dots, w_n)| = \sqrt{\det(g_{ij})},$$

where $g_{ij} = T(w_i, w_j)$. *Hint:* If v_1, \dots, v_n is an orthonormal basis and $w_i = \sum_{j=1}^n a_{ij}v_j$, show that $g_{ij} = \sum_{k=1}^n a_{ik}a_{kj}$.

- 4-4. If ω is the volume element of V determined by T and μ , and $f: \mathbf{R}^n \rightarrow V$ is an isomorphism such that $f^*T = \langle, \rangle$ and such that $[f(e_1), \dots, f(e_n)] = \mu$, show that $f^*\omega = \det$.

- 4-5. If $c: [0,1] \rightarrow (\mathbf{R}^n)^n$ is continuous and each $(c^1(t), \dots, c^n(t))$ is a basis for \mathbf{R}^n , show that $[c^1(0), \dots, c^n(0)] = [c^1(1), \dots, c^n(1)]$.

Hint: Consider $\det \circ c$.

- 4-6. (a) If $v \in \mathbf{R}^2$, what is $v \times$?

(b) If $v_1, \dots, v_{n-1} \in \mathbf{R}^n$ are linearly independent, show that $[v_1, \dots, v_{n-1}, v_1 \times \dots \times v_{n-1}]$ is the usual orientation of \mathbf{R}^n .

- 4-7. Show that every non-zero $\omega \in \Lambda^n(V)$ is the volume element determined by some inner product T and orientation μ for V .

- 4-8. If $\omega \in \Lambda^n(V)$ is a volume element, define a "cross product" $v_1 \times \dots \times v_{n-1}$ in terms of ω .

- 4-9.* Deduce the following properties of the cross product in \mathbf{R}^3 :

$$\begin{array}{lll} \text{(a)} & e_1 \times e_1 = 0 & e_2 \times e_1 = -e_3 & e_3 \times e_1 = e_2 \\ & e_1 \times e_2 = e_3 & e_2 \times e_2 = 0 & e_3 \times e_2 = -e_1 \\ & e_1 \times e_3 = -e_2 & e_2 \times e_3 = e_1 & e_3 \times e_3 = 0. \end{array}$$

$$\begin{aligned} \text{(b)} \quad v \times w &= (v^2w^3 - v^3w^2)e_1 \\ &\quad + (v^3w^1 - v^1w^3)e_2 \\ &\quad + (v^1w^2 - v^2w^1)e_3. \end{aligned}$$

$$\text{(c)} \quad |v \times w| = |v| \cdot |w| \cdot |\sin \theta|, \text{ where } \theta = \angle(v, w).$$

$$\langle v \times w, v \rangle = \langle v \times w, w \rangle = 0.$$

$$\text{(d)} \quad \langle v, w \times z \rangle = \langle w, z \times v \rangle = \langle z, v \times w \rangle$$

$$v \times (w \times z) = \langle v, z \rangle w - \langle v, w \rangle z$$

$$(v \times w) \times z = \langle v, z \rangle w - \langle w, z \rangle v.$$

$$\text{(e)} \quad |v \times w| = \sqrt{\langle v, v \rangle \cdot \langle w, w \rangle - \langle v, w \rangle^2}.$$

- 4-10. If $w_1, \dots, w_{n-1} \in \mathbf{R}^n$, show that

$$|w_1 \times \dots \times w_{n-1}| = \sqrt{\det(g_{ij})},$$

where $g_{ij} = \langle w_i, w_j \rangle$. *Hint:* Apply Problem 4-3 to a certain $(n-1)$ -dimensional subspace of \mathbf{R}^n .

- 4-11. If T is an inner product on V , a linear transformation $f: V \rightarrow V$ is called **self-adjoint** (with respect to T) if $T(x, f(y)) = T(f(x), y)$ for $x, y \in V$. If v_1, \dots, v_n is an orthonormal basis and $A = (a_{ij})$ is the matrix of f with respect to this basis, show that $a_{ij} = a_{ji}$.

- 4-12. If $f_1, \dots, f_{n-1}: \mathbf{R}^m \rightarrow \mathbf{R}^n$, define $f_1 \times \dots \times f_{n-1}: \mathbf{R}^m \rightarrow \mathbf{R}^n$ by $f_1 \times \dots \times f_{n-1}(p) = f_1(p) \times \dots \times f_{n-1}(p)$. Use Problem 2-14 to derive a formula for $D(f_1 \times \dots \times f_{n-1})$ when f_1, \dots, f_{n-1} are differentiable.

FIELDS AND FORMS

If $p \in \mathbf{R}^n$, the set of all pairs (p, v) , for $v \in \mathbf{R}^n$, is denoted \mathbf{R}^n_p , and called the **tangent space** of \mathbf{R}^n at p . This set is made into a vector space in the most obvious way, by defining

$$\begin{aligned}(p, v) + (p, w) &= (p, v + w), \\ a \cdot (p, v) &= (p, av).\end{aligned}$$

A vector $v \in \mathbf{R}^n$ is often pictured as an arrow from 0 to v ; the vector $(p, v) \in \mathbf{R}^n_p$ may be pictured (Figure 4-1) as an arrow with the same direction and length, but with initial point p . This arrow goes from p to the point $p + v$, and we therefore

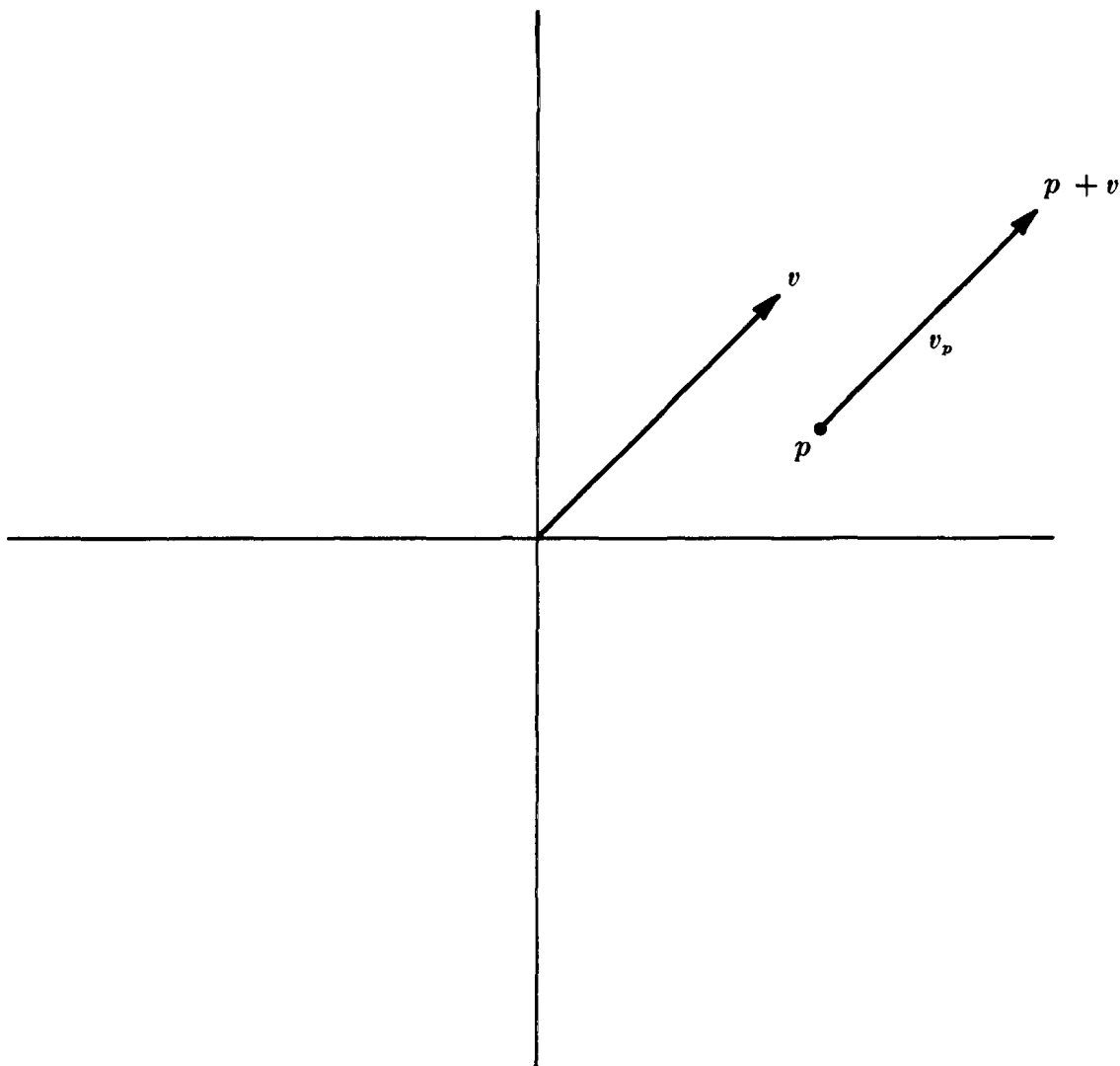


FIGURE 4-1

define $p + v$ to be the **end point** of (p, v) . We will usually write (p, v) as v_p (read: the vector v at p).

The vector space \mathbf{R}^n_p is so closely allied to \mathbf{R}^n that many of the structures on \mathbf{R}^n have analogues on \mathbf{R}^n_p . In particular the **usual inner product** $\langle \cdot, \cdot \rangle_p$ for \mathbf{R}^n_p is defined by $\langle v_p, w_p \rangle_p = \langle v, w \rangle$, and the **usual orientation** for \mathbf{R}^n_p is $[(e_1)_p, \dots, (e_n)_p]$.

Any operation which is possible in a vector space may be performed in each \mathbf{R}^n_p , and most of this section is merely an elaboration of this theme. About the simplest operation in a vector space is the selection of a vector from it. If such a selection is made in each \mathbf{R}^n_p , we obtain a **vector field** (Figure 4-2). To be precise, a vector field is a function F such that $F(p) \in \mathbf{R}^n_p$ for each $p \in \mathbf{R}^n$. For each p there are numbers $F^1(p), \dots, F^n(p)$ such that

$$F(p) = F^1(p) \cdot (e_1)_p + \dots + F^n(p) \cdot (e_n)_p.$$

We thus obtain n **component functions** $F^i: \mathbf{R}^n \rightarrow \mathbf{R}$. The vector field F is called continuous, differentiable, etc., if the functions F^i are. Similar definitions can be made for a vector field defined only on an open subset of \mathbf{R}^n . Operations on vectors yield operations on vector fields when applied at each point separately. For example, if F and G are vector fields

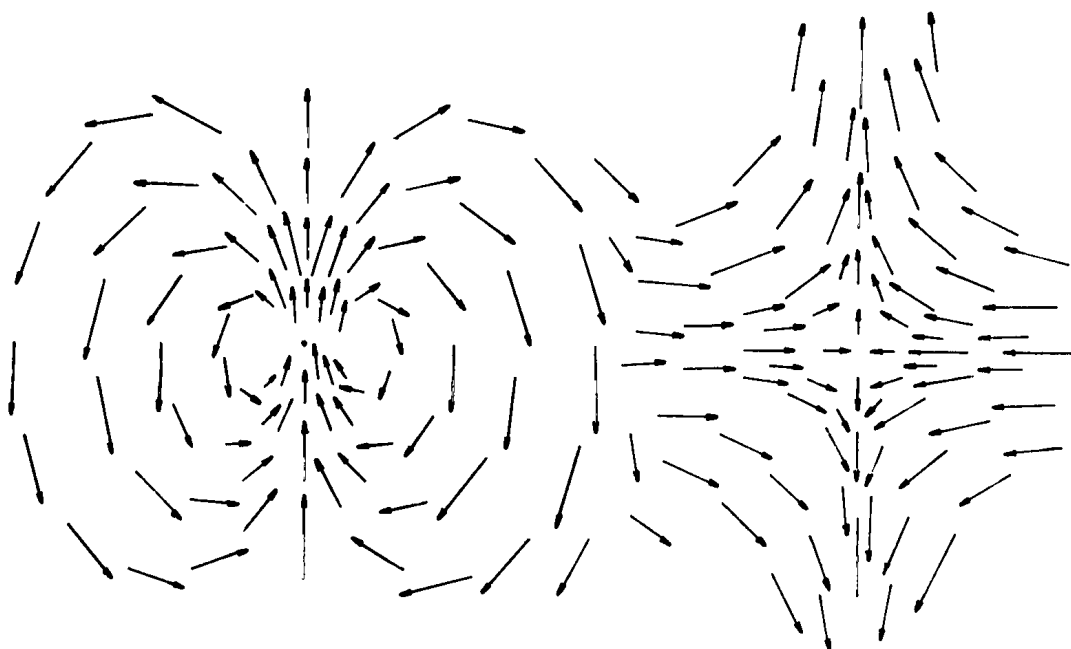


FIGURE 4-2

and f is a function, we define

$$\begin{aligned}(F + G)(p) &= F(p) + G(p), \\ \langle F, G \rangle(p) &= \langle F(p), G(p) \rangle, \\ (f \cdot F)(p) &= f(p)F(p).\end{aligned}$$

If F_1, \dots, F_{n-1} are vector fields on \mathbf{R}^n , then we can similarly define

$$(F_1 \times \dots \times F_{n-1})(p) = F_1(p) \times \dots \times F_{n-1}(p).$$

Certain other definitions are standard and useful. We define the **divergence**, $\operatorname{div} F$ of F , as $\sum_{i=1}^n D_i F^i$. If we introduce the formal symbolism

$$\nabla = \sum_{i=1}^n D_i \cdot e_i,$$

we can write, symbolically, $\operatorname{div} F = \langle \nabla, F \rangle$. If $n = 3$ we write, in conformity with this symbolism,

$$\begin{aligned}(\nabla \times F)(p) &= (D_2 F^3 - D_3 F^2)(e_1)_p \\ &\quad + (D_3 F^1 - D_1 F^3)(e_2)_p \\ &\quad + (D_1 F^2 - D_2 F^1)(e_3)_p.\end{aligned}$$

The vector field $\nabla \times F$ is called $\operatorname{curl} F$. The names “divergence” and “curl” are derived from physical considerations which are explained at the end of this book.

Many similar considerations may be applied to a function ω with $\omega(p) \in \Lambda^k(\mathbf{R}^n_p)$; such a function is called a **k -form** on \mathbf{R}^n , or simply a **differential form**. If $\varphi_1(p), \dots, \varphi_n(p)$ is the dual basis to $(e_1)_p, \dots, (e_n)_p$, then

$$\omega(p) = \sum_{i_1 < \dots < i_k} \omega_{i_1, \dots, i_k}(p) \cdot [\varphi_{i_1}(p) \wedge \dots \wedge \varphi_{i_k}(p)]$$

for certain functions ω_{i_1, \dots, i_k} ; the form ω is called continuous, differentiable, etc., if these functions are. We shall usually assume tacitly that forms and vector fields are differentiable, and “differentiable” will henceforth mean “ C^∞ ”; this is a simplifying assumption that eliminates the need for counting how many times a function is differentiated in a proof. The sum $\omega + \eta$, product $f \cdot \omega$, and wedge product $\omega \wedge \eta$ are defined

in the obvious way. A function f is considered to be a 0-form and $f \cdot \omega$ is also written $f \wedge \omega$.

If $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable, then $Df(p) \in \Lambda^1(\mathbf{R}^n)$. By a minor modification we therefore obtain a 1-form df , defined by

$$df(p)(v_p) = Df(p)(v).$$

Let us consider in particular the 1-forms $d\pi^i$. It is customary to let x^i denote the function π^i . (On \mathbf{R}^3 we often denote x^1 , x^2 , and x^3 by x , y , and z .) This standard notation has obvious disadvantages but it allows many classical results to be expressed by formulas of equally classical appearance. Since $dx^i(p)(v_p) = d\pi^i(p)(v_p) = D\pi^i(p)(v) = v^i$, we see that $dx^1(p), \dots, dx^n(p)$ is just the dual basis to $(e_1)_p, \dots, (e_n)_p$. Thus every k -form ω can be written

$$\omega = \sum_{i_1 < \dots < i_k} \omega_{i_1, \dots, i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k}.$$

The expression for df is of particular interest.

4-7 Theorem. *If $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable, then*

$$df = D_1 f \cdot dx^1 + \dots + D_n f \cdot dx^n.$$

In classical notation,

$$df = \frac{\partial f}{\partial x^1} dx^1 + \dots + \frac{\partial f}{\partial x^n} dx^n.$$

Proof. $df(p)(v_p) = Df(p)(v) = \sum_{i=1}^n v^i \cdot D_i f(p)$
 $= \sum_{i=1}^n dx^i(p)(v_p) \cdot D_i f(p). \quad \blacksquare$

If we consider now a differentiable function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ we have a linear transformation $Df(p): \mathbf{R}^n \rightarrow \mathbf{R}^m$. Another minor modification therefore produces a linear transformation $f_*: \mathbf{R}^n_p \rightarrow \mathbf{R}^m_{f(p)}$ defined by

$$f_*(v_p) \doteq (Df(p)(v))_{f(p)}.$$

This linear transformation induces a linear transformation $f^*: \Lambda^k(\mathbf{R}^m_{f(p)}) \rightarrow \Lambda^k(\mathbf{R}^n_p)$. If ω is a k -form on \mathbf{R}^m we can therefore define a k -form $f^*\omega$ on \mathbf{R}^n by $(f^*\omega)(p) = f^*(\omega(f(p)))$.

Recall this means that if $v_1, \dots, v_k \in \mathbf{R}^n_p$, then we have $f^*\omega(p)(v_1, \dots, v_k) = \omega(f(p))(f_*(v_1), \dots, f_*(v_k))$. As an antidote to the abstractness of these definitions we present a theorem, summarizing the important properties of f^* , which allows explicit calculations of $f^*\omega$.

4-8 Theorem. *If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable, then*

$$(1) \quad f^*(dx^i) = \sum_{j=1}^n D_j f^i \cdot dx^j = \sum_{j=1}^n \frac{\partial f^i}{\partial x^j} dx^j.$$

$$(2) \quad f^*(\omega_1 + \omega_2) = f^*(\omega_1) + f^*(\omega_2).$$

$$(3) \quad f^*(g \cdot \omega) = (g \circ f) \cdot f^*\omega.$$

$$(4) \quad f^*(\omega \wedge \eta) = f^*\omega \wedge f^*\eta.$$

Proof

$$\begin{aligned} (1) \quad f^*(dx^i)(p)(v_p) &= dx^i(f(p))(f_*v_p) \\ &= dx^i(f(p))(\sum_{j=1}^n v^j \cdot D_j f^i(p), \dots, \sum_{j=1}^n v^j \cdot D_j f^m(p))_{f(p)} \\ &= \sum_{j=1}^n v^j \cdot D_j f^i(p) \\ &= \sum_{j=1}^n D_j f^i(p) \cdot dx^j(p)(v_p). \end{aligned}$$

The proofs of (2), (3), and (4) are left to the reader. ■

By repeatedly applying Theorem 4-8 we have, for example,

$$\begin{aligned} f^*(P dx^1 \wedge dx^2 + Q dx^2 \wedge dx^3) &= (P \circ f)[f^*(dx^1) \wedge f^*(dx^2)] \\ &\quad + (Q \circ f)[f^*(dx^2) \wedge f^*(dx^3)]. \end{aligned}$$

The expression obtained by expanding out each $f^*(dx^i)$ is quite complicated. (It is helpful to remember, however, that we have $dx^i \wedge dx^i = (-1)dx^i \wedge dx^i = 0$.) In one special case it will be worth our while to make an explicit evaluation.

4-9 Theorem. *If $f: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is differentiable, then*

$$f^*(h dx^1 \wedge \dots \wedge dx^n) = (h \circ f)(\det f') dx^1 \wedge \dots \wedge dx^n.$$

Proof. Since

$$f^*(h dx^1 \wedge \dots \wedge dx^n) = (h \circ f)f^*(dx^1 \wedge \dots \wedge dx^n),$$

it suffices to show that

$$f^*(dx^1 \wedge \cdots \wedge dx^n) = (\det f') dx^1 \wedge \cdots \wedge dx^n.$$

Let $p \in \mathbf{R}^n$ and let $A = (a_{ij})$ be the matrix of $f'(p)$. Here, and whenever convenient and not confusing, we shall omit “ p ” in $dx^1 \wedge \cdots \wedge dx^n(p)$, etc. Then

$$\begin{aligned} f^*(dx^1 \wedge \cdots \wedge dx^n)(e_1, \dots, e_n) &= dx^1 \wedge \cdots \wedge dx^n(f_*e_1, \dots, f_*e_n) \\ &= dx^1 \wedge \cdots \wedge \left(\sum_{i=1}^n a_{i1}e_i, \dots, \sum_{i=1}^n a_{in}e_i \right) \\ &= \det(a_{ij}) \cdot dx^1 \wedge \cdots \wedge dx^n(e_1, \dots, e_n), \end{aligned}$$

by Theorem 4-6. ■

An important construction associated with forms is a generalization of the operator d which changes 0-forms into 1-forms. If

$$\omega = \sum_{i_1 < \cdots < i_k} \omega_{i_1, \dots, i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k},$$

we define a $(k+1)$ -form $d\omega$, the **differential** of ω , by

$$\begin{aligned} d\omega &= \sum_{i_1 < \cdots < i_k} d\omega_{i_1, \dots, i_k} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k} \\ &= \sum_{i_1 < \cdots < i_k} \sum_{\alpha=1}^n D_\alpha(\omega_{i_1, \dots, i_k}) \cdot dx^\alpha \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}. \end{aligned}$$

4-10 Theorem

- (1) $d(\omega + \eta) = d\omega + d\eta$.
- (2) If ω is a k -form and η is an l -form, then

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^k \omega \wedge d\eta.$$

- (3) $d(d\omega) = 0$. Briefly, $d^2 = 0$.
- (4) If ω is a k -form on \mathbf{R}^m and $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is differentiable, then $f^*(d\omega) = d(f^*\omega)$.

Proof

- (1) Left to the reader.
- (2) The formula is true if $\omega = dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ and $\eta = dx^{j_1} \wedge \cdots \wedge dx^{j_l}$, since all terms vanish. The formula is easily checked when ω is a 0-form. The general formula may be derived from (1) and these two observations.
- (3) Since

$$d\omega = \sum_{i_1 < \cdots < i_k} \sum_{\alpha=1}^n D_{\alpha}(\omega_{i_1, \dots, i_k}) dx^{\alpha} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k},$$

we have

$$d(d\omega) = \sum_{i_1 < \cdots < i_k} \sum_{\alpha=1}^n \sum_{\beta=1}^n D_{\alpha, \beta}(\omega_{i_1, \dots, i_k}) dx^{\beta} \wedge dx^{\alpha} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}.$$

In this sum the terms

$$D_{\alpha, \beta}(\omega_{i_1, \dots, i_k}) dx^{\beta} \wedge dx^{\alpha} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}$$

and

$$D_{\beta, \alpha}(\omega_{i_1, \dots, i_k}) dx^{\alpha} \wedge dx^{\beta} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}$$

cancel in pairs.

- (4) This is clear if ω is a 0-form. Suppose, inductively, that (4) is true when ω is a k -form. It suffices to prove (4) for a $(k+1)$ -form of the type $\omega \wedge dx^i$. We have

$$\begin{aligned} f^*(d(\omega \wedge dx^i)) &= f^*(d\omega \wedge dx^i + (-1)^k \omega \wedge d(dx^i)) \\ &= f^*(d\omega \wedge dx^i) = f^*(d\omega) \wedge f^*(dx^i) \\ &= d(f^*\omega \wedge f^*(dx^i)) \quad \text{by (2) and (3)} \\ &= d(f^*(\omega \wedge dx^i)). \quad \blacksquare \end{aligned}$$

A form ω is called **closed** if $d\omega = 0$ and **exact** if $\omega = d\eta$, for some η . Theorem 4-10 shows that every exact form is closed, and it is natural to ask whether, conversely, every closed form is exact. If ω is the 1-form $P dx + Q dy$ on \mathbf{R}^2 , then

$$\begin{aligned} d\omega &= (D_1P dx + D_2P dy) \wedge dx + (D_1Q dx + D_2Q dy) \wedge dy \\ &= (D_1Q - D_2P) dx \wedge dy. \end{aligned}$$

Thus, if $d\omega = 0$, then $D_1Q = D_2P$. Problems 2-21 and 3-34 show that there is a 0-form f such that $\omega = df = D_1f dx + D_2f dy$. If ω is defined only on a subset of \mathbf{R}^2 , however, such a function may not exist. The classical example is the form

$$\omega = \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy$$

defined on $\mathbf{R}^2 - 0$. This form is usually denoted $d\theta$ (where θ is defined in Problem 3-41), since (Problem 4-21) it equals $d\theta$ on the set $\{(x,y): x < 0, \text{ or } x \geq 0 \text{ and } y \neq 0\}$, where θ is defined. Note, however, that θ cannot be defined continuously on all of $\mathbf{R}^2 - 0$. If $\omega = df$ for some function $f: \mathbf{R}^2 - 0 \rightarrow \mathbf{R}$, then $D_1f = D_1\theta$ and $D_2f = D_2\theta$, so $f = \theta + \text{constant}$, showing that such an f cannot exist.

Suppose that $\omega = \sum_{i=1}^n \omega_i dx^i$ is a 1-form on \mathbf{R}^n and ω happens to equal $df = \sum_{i=1}^n D_i f \cdot dx^i$. We can clearly assume that $f(0) = 0$. As in Problem 2-35, we have

$$\begin{aligned} f(x) &= \int_0^1 \frac{d}{dt} f(tx) dt \\ &= \int_0^1 \sum_{i=1}^n D_i f(tx) \cdot x^i dt \\ &= \int_0^1 \sum_{i=1}^n \omega_i(tx) \cdot x^i dt. \end{aligned}$$

This suggests that in order to find f , given ω , we consider the function $I\omega$, defined by

$$I\omega(x) = \int_0^1 \sum_{i=1}^n \omega_i(tx) \cdot x^i dt.$$

Note that the definition of $I\omega$ makes sense if ω is defined only on an open set $A \subset \mathbf{R}^n$ with the property that whenever $x \in A$, the line segment from 0 to x is contained in A ; such an open set is called **star-shaped** with respect to 0 (Figure 4-3). A somewhat involved calculation shows that (on a star-shaped open set) we have $\omega = d(I\omega)$ provided that ω satisfies the necessary condition $d\omega = 0$. The calculation, as well as the definition of $I\omega$, may be generalized considerably:

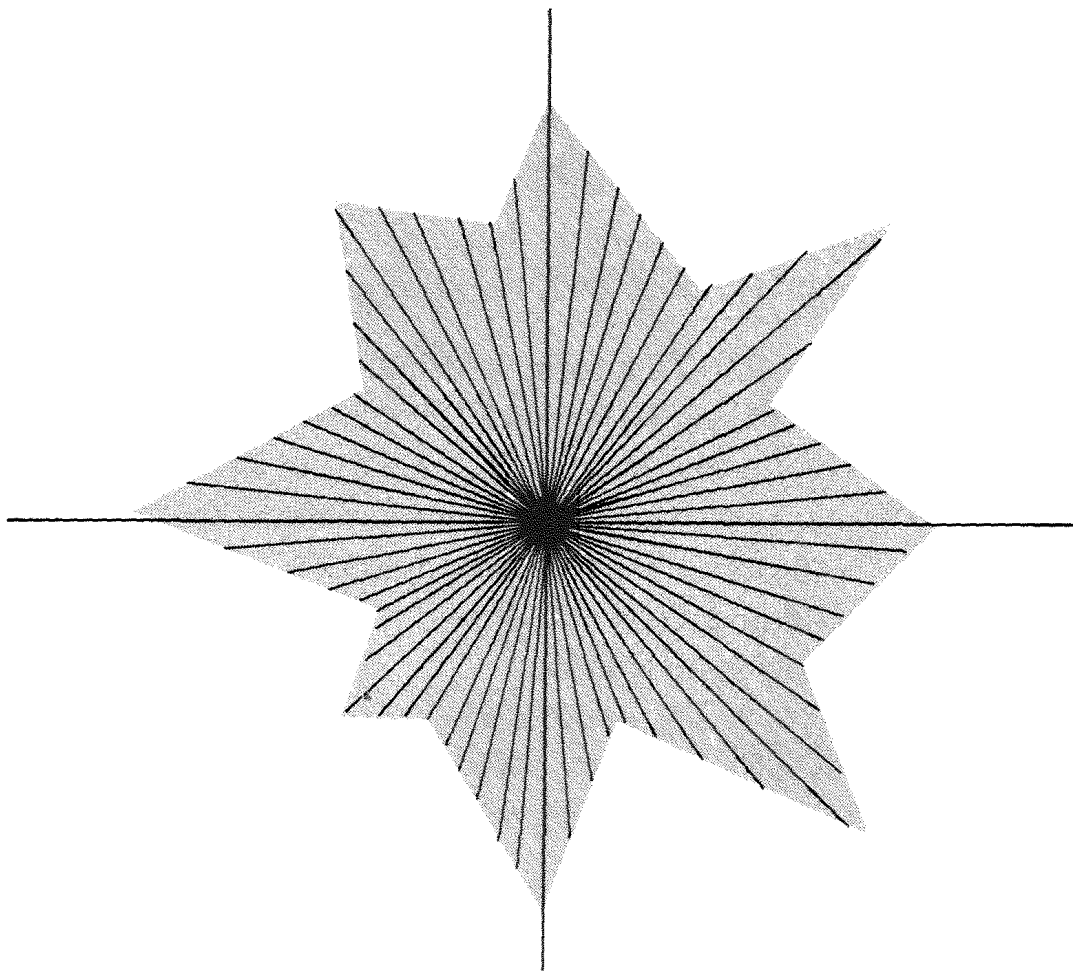


FIGURE 4-3

4-11 Theorem (Poincaré Lemma). *If $A \subset \mathbf{R}^n$ is an open set star-shaped with respect to 0, then every closed form on A is exact.*

Proof. We will define a function I from l -forms to $(l-1)$ -forms (for each l), such that $I(0) = 0$ and $\omega = I(d\omega) + d(I\omega)$ for any form ω . It follows that $\omega = d(I\omega)$ if $d\omega = 0$. Let

$$\omega = \sum_{i_1 < \dots < i_l} \omega_{i_1, \dots, i_l} dx^{i_1} \wedge \dots \wedge dx^{i_l}.$$

Since A is star-shaped we can define

$$I\omega(x) = \sum_{i_1 < \dots < i_l} \sum_{\alpha=1}^l (-1)^{\alpha-1} \left(\int_0^1 t^{l-1} \omega_{i_1, \dots, i_l}(tx) dt \right) x^{i_\alpha} dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\alpha}} \wedge \dots \wedge dx^{i_l}.$$

(The symbol \wedge over dx^{i_α} indicates that it is omitted.) The

proof that $\omega = I(d\omega) + d(I\omega)$ is an elaborate computation: We have, using Problem 3-32,

$$\begin{aligned} d(I\omega) &= l \cdot \sum_{i_1 < \dots < i_l} \left(\int_0^1 t^{l-1} \omega_{i_1, \dots, i_l}(tx) dt \right) \\ &\quad dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &+ \sum_{i_1 < \dots < i_l} \sum_{\alpha=1}^l \sum_{j=1}^n (-1)^{\alpha-1} \left(\int_0^1 t^l D_j(\omega_{i_1, \dots, i_l})(tx) dt \right) x^{i_\alpha} \\ &\quad dx^j \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\alpha}} \wedge \dots \wedge dx^{i_l}. \end{aligned}$$

(Explain why we have the factor t^l , instead of t^{l-1} .) We also have

$$d\omega = \sum_{i_1 < \dots < i_l} \sum_{j=1}^n D_j(\omega_{i_1, \dots, i_l}) \cdot dx^j \wedge dx^{i_1} \wedge \dots \wedge dx^{i_l}.$$

Applying I to the $(l+1)$ -form $d\omega$, we obtain

$$\begin{aligned} I(d\omega) &= \sum_{i_1 < \dots < i_l} \sum_{j=1}^n \left(\int_0^1 t^l D_j(\omega_{i_1, \dots, i_l})(tx) dt \right) x^j \\ &\quad dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &- \sum_{i_1 < \dots < i_l} \sum_{j=1}^n \sum_{\alpha=1}^l (-1)^{\alpha-1} \left(\int_0^1 t^l D_j(\omega_{i_1, \dots, i_l})(tx) dt \right) x^{i_\alpha} \\ &\quad dx^j \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\alpha}} \wedge \dots \wedge dx^{i_l}. \end{aligned}$$

Adding, the triple sums cancel, and we obtain

$$\begin{aligned} d(I\omega) + I(d\omega) &= \sum_{i_1 < \dots < i_l} l \cdot \left(\int_0^1 t^{l-1} \omega_{i_1, \dots, i_l}(tx) dt \right) \\ &\quad dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &+ \sum_{i_1 < \dots < i_l} \sum_{j=1}^n \left(\int_0^1 t^l x^j D_j(\omega_{i_1, \dots, i_l})(tx) dt \right) \\ &\quad dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &= \sum_{i_1 < \dots < i_l} \left(\int_0^1 \frac{d}{dt} [t^l \omega_{i_1, \dots, i_l}(tx)] dt \right) \\ &\quad dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &= \sum_{i_1 < \dots < i_l} \omega_{i_1, \dots, i_l} dx^{i_1} \wedge \dots \wedge dx^{i_l} \\ &= \omega. \quad \blacksquare \end{aligned}$$

Problems. 4-13. (a) If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ and $g: \mathbf{R}^m \rightarrow \mathbf{R}^p$, show that $(g \circ f)_* = g_* \circ f_*$ and $(g \circ f)^* = f^* \circ g^*$.

(b) If $f, g: \mathbf{R}^n \rightarrow \mathbf{R}$, show that $d(f \cdot g) = f \cdot dg + g \cdot df$.

4-14. Let c be a differentiable curve in \mathbf{R}^n , that is, a differentiable function $c: [0, 1] \rightarrow \mathbf{R}^n$. Define the **tangent vector** v of c at t as $c_*((e_1)_t) = ((c^1)'(t), \dots, (c^n)'(t))_{c(t)}$. If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, show that the tangent vector to $f \circ c$ at t is $f_*(v)$.

4-15. Let $f: \mathbf{R} \rightarrow \mathbf{R}$ and define $c: \mathbf{R} \rightarrow \mathbf{R}^2$ by $c(t) = (t, f(t))$. Show that the end point of the tangent vector of c at t lies on the tangent line to the graph of f at $(t, f(t))$.

4-16. Let $c: [0, 1] \rightarrow \mathbf{R}^n$ be a curve such that $|c(t)| = 1$ for all t . Show that $c(t)_{c(t)}$ and the tangent vector to c at t are perpendicular.

4-17. If $f: \mathbf{R}^n \rightarrow \mathbf{R}^n$, define a vector field \mathbf{f} by $\mathbf{f}(p) = f(p)_p \in \mathbf{R}^n_p$.

(a) Show that every vector field F on \mathbf{R}^n is of the form \mathbf{f} for some f .

(b) Show that $\operatorname{div} \mathbf{f} = \operatorname{trace} f'$.

4-18. If $f: \mathbf{R}^n \rightarrow \mathbf{R}$, define a vector field $\operatorname{grad} f$ by

$$(\operatorname{grad} f)(p) = D_1 f(p) \cdot (e_1)_p + \dots + D_n f(p) \cdot (e_n)_p.$$

For obvious reasons we also write $\operatorname{grad} f = \nabla f$. If $\nabla f(p) = w_p$, prove that $D_v f(p) = \langle v, w \rangle$ and conclude that $\nabla f(p)$ is the direction in which f is changing fastest at p .

4-19. If F is a vector field on \mathbf{R}^3 , define the forms

$$\begin{aligned}\omega^1_F &= F^1 dx + F^2 dy + F^3 dz, \\ \omega^2_F &= F^1 dy \wedge dz + F^2 dz \wedge dx + F^3 dx \wedge dy.\end{aligned}$$

(a) Prove that

$$\begin{aligned}df &= \omega^1_{\operatorname{grad} f}, \\ d(\omega^1_F) &= \omega^2_{\operatorname{curl} F}, \\ d(\omega^2_F) &= (\operatorname{div} F) dx \wedge dy \wedge dz.\end{aligned}$$

(b) Use (a) to prove that

$$\begin{aligned}\operatorname{curl} \operatorname{grad} f &= 0, \\ \operatorname{div} \operatorname{curl} F &= 0.\end{aligned}$$

(c) If F is a vector field on a star-shaped open set A and $\operatorname{curl} F = 0$, show that $F = \operatorname{grad} f$ for some function $f: A \rightarrow \mathbf{R}$. Similarly, if $\operatorname{div} F = 0$, show that $F = \operatorname{curl} G$ for some vector field G on A .

4-20. Let $f: U \rightarrow \mathbf{R}^n$ be a differentiable function with a differentiable inverse $f^{-1}: f(U) \rightarrow \mathbf{R}^n$. If every closed form on U is exact, show that the same is true for $f(U)$. *Hint:* If $d\omega = 0$ and $f^*\omega = d\eta$, consider $(f^{-1})^*\eta$.

4-21.* Prove that on the set where θ is defined we have

$$d\theta = \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy.$$

GEOMETRIC PRELIMINARIES

A **singular n -cube** in $A \subset \mathbf{R}^n$ is a continuous function $c: [0,1]^n \rightarrow A$ (here $[0,1]^n$ denotes the n -fold product $[0,1] \times \cdots \times [0,1]$). We let \mathbf{R}^0 and $[0,1]^0$ both denote $\{0\}$. A singular 0-cube in A is then a function $f: \{0\} \rightarrow A$ or, what amounts to the same thing, a point in A . A singular 1-cube is often called a **curve**. A particularly simple, but particularly important example of a singular n -cube in \mathbf{R}^n is the **standard n -cube** $I^n: [0,1]^n \rightarrow \mathbf{R}^n$ defined by $I^n(x) = x$ for $x \in [0,1]^n$.

We shall need to consider formal sums of singular n -cubes in A multiplied by integers, that is, expressions like

$$2c_1 + 3c_2 - 4c_3,$$

where c_1, c_2, c_3 are singular n -cubes in A . Such a finite sum of singular n -cubes with integer coefficients is called an **n -chain** in A . In particular a singular n -cube c is also considered as an n -chain $1 \cdot c$. It is clear how n -chains can be added, and multiplied by integers. For example

$$2(c_1 + 3c_4) + (-2)(c_1 + c_3 + c_2) = -2c_2 - 2c_3 + 6c_4.$$

(A rigorous exposition of this formalism is presented in Problem 4-22.)

For each singular n -chain c in A we shall define an $(n-1)$ -chain in A called the **boundary** of c and denoted ∂c . The boundary of I^2 , for example, might be defined as the sum of four singular 1-cubes arranged counterclockwise around the boundary of $[0,1]^2$, as indicated in Figure 4-4(a). It is actually much more convenient to define ∂I^2 as the sum, with the indicated coefficients, of the four singular 1-cubes shown in Figure 4-4(b). The precise definition of ∂I^n requires some preliminary notions. For each i with $1 \leq i \leq n$ we define two singular $(n-1)$ -cubes $I_{(i,0)}^n$ and $I_{(i,1)}^n$ as follows. If

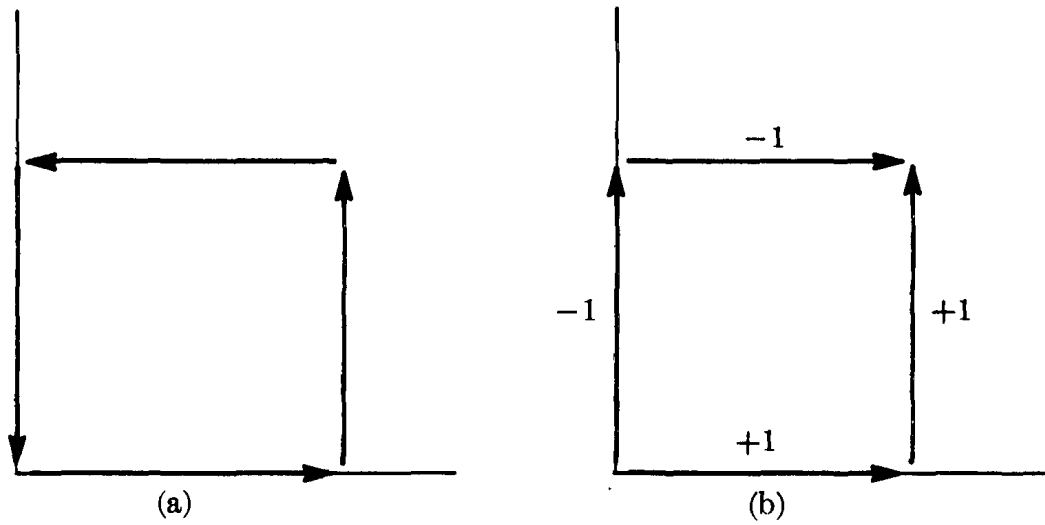


FIGURE 4-4

$x \in [0,1]^{n-1}$, then

$$\begin{aligned} I_{(i,0)}^n(x) &= I^n(x^1, \dots, x^{i-1}, 0, x^i, \dots, x^{n-1}) \\ &= (x^1, \dots, x^{i-1}, 0, x^i, \dots, x^{n-1}), \\ I_{(i,1)}^n(x) &= I^n(x^1, \dots, x^{i-1}, 1, x^i, \dots, x^{n-1}) \\ &= (x^1, \dots, x^{i-1}, 1, x^i, \dots, x^{n-1}). \end{aligned}$$

We call $I_{(i,0)}^n$ the $(i,0)$ -face of I^n and $I_{(i,1)}^n$ the $(i,1)$ -face (Figure 4-5). We then define

$$\partial I^n = \sum_{i=1}^n \sum_{\alpha=0,1} (-1)^{i+\alpha} I_{(i,\alpha)}^n.$$

For a general singular n -cube $c: [0,1]^n \rightarrow A$ we first define the (i,α) -face,

$$c_{(i,\alpha)} = c \circ (I_{(i,\alpha)}^n)$$

and then define

$$\partial c = \sum_{i=1}^n \sum_{\alpha=0,1} (-1)^{i+\alpha} c_{(i,\alpha)}.$$

Finally we define the boundary of an n -chain $\sum a_i c_i$ by

$$\partial(\sum a_i c_i) = \sum a_i \partial(c_i).$$

Although these few definitions suffice for all applications in this book, we include here the one standard property of ∂ .

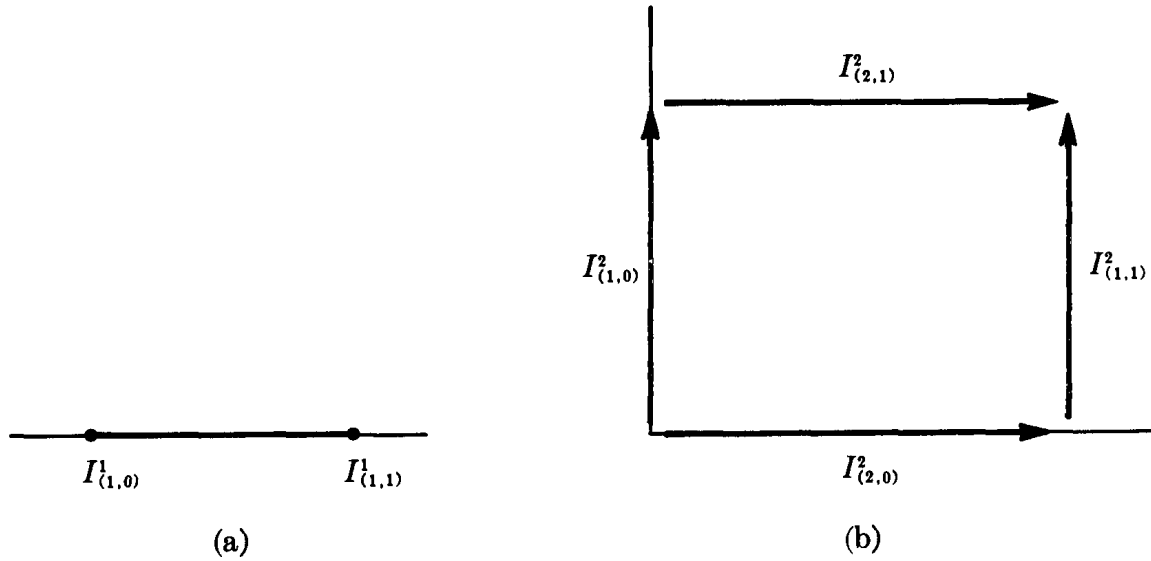


FIGURE 4-5

4-12 Theorem. If c is an n -chain in A , then $\partial(\partial c) = 0$. Briefly, $\partial^2 = 0$.

Proof. Let $i \leq j$ and consider $(I_{(i,\alpha)}^n)_{(j,\beta)}$. If $x \in [0,1]^{n-2}$, then, remembering the definition of the (j,β) -face of a singular n -cube, we have

$$\begin{aligned} (I_{(i,\alpha)}^n)_{(j,\beta)}(x) &= I_{(i,\alpha)}^n(I_{(j,\beta)}^{n-1}(x)) \\ &= I_{(i,\alpha)}^n(x^1, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}) \\ &= I^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}). \end{aligned}$$

Similarly

$$\begin{aligned} (I_{(j+1,\beta)}^n)_{(i,\alpha)} &= I_{(j+1,\beta)}^n(I_{(i,\alpha)}^{n-1}(x)) \\ &= I_{(j+1,\beta)}^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{n-2}) \\ &= I^n(x^1, \dots, x^{i-1}, \alpha, x^i, \dots, x^{j-1}, \beta, x^j, \dots, x^{n-2}). \end{aligned}$$

Thus $(I_{(i,\alpha)}^n)_{(j,\beta)} = (I_{(j+1,\beta)}^n)_{(i,\alpha)}$ for $i \leq j$. (It may help to verify this in Figure 4-5.) It follows easily for any singular n -cube c that $(c_{(i,\alpha)})_{(j,\beta)} = (c_{(j+1,\beta)})_{(i,\alpha)}$ when $i \leq j$. Now

$$\begin{aligned} \partial(\partial c) &= \partial \left(\sum_{i=1}^n \sum_{\alpha=0,1} (-1)^{i+\alpha} c_{(i,\alpha)} \right) \\ &= \sum_{i=1}^n \sum_{\alpha=0,1} \sum_{j=1}^{n-1} \sum_{\beta=0,1} (-1)^{i+\alpha+j+\beta} (c_{(i,\alpha)})_{(j,\beta)}. \end{aligned}$$

In this sum $(c_{(i,\alpha)})_{(j,\beta)}$ and $(c_{(j+1,\beta)})_{(i,\alpha)}$ occur with opposite signs. Therefore all terms cancel out in pairs and $\partial(\partial c) = 0$. Since the theorem is true for any singular n -cube, it is also true for singular n -chains. ■

It is natural to ask whether Theorem 4-12 has a converse: If $\partial c = 0$, is there a chain d in A such that $c = \partial d$? The answer depends on A and is generally "no." For example, define $c: [0,1] \rightarrow \mathbb{R}^2 - 0$ by $c(t) = (\sin 2\pi nt, \cos 2\pi nt)$, where n is a non-zero integer. Then $c(1) = c(0)$, so $\partial c = 0$. But (Problem 4-26) there is no 2-chain c' in $\mathbb{R}^2 - 0$, with $\partial c' = c$.

Problems. 4-22. Let \mathcal{S} be the set of all singular n -cubes, and \mathbb{Z} the integers. An n -chain is a function $f: \mathcal{S} \rightarrow \mathbb{Z}$ such that $f(c) = 0$ for all but finitely many c . Define $f + g$ and nf by $(f + g)(c) = f(c) + g(c)$ and $nf(c) = n \cdot f(c)$. Show that $f + g$ and nf are n -chains if f and g are. If $c \in \mathcal{S}$, let c also denote the function f such that $f(c) = 1$ and $f(c') = 0$ for $c' \neq c$. Show that every n -chain f can be written $a_1 c_1 + \cdots + a_k c_k$ for some integers a_1, \dots, a_k and singular n -cubes c_1, \dots, c_k .

4-23. For $R > 0$ and n an integer, define the singular 1-cube $c_{R,n}: [0,1] \rightarrow \mathbb{R}^2 - 0$ by $c_{R,n}(t) = (R \cos 2\pi nt, R \sin 2\pi nt)$. Show that there is a singular 2-cube $c: [0,1]^2 \rightarrow \mathbb{R}^2 - 0$ such that $c_{R_1,n} - c_{R_2,n} = \partial c$.

4-24. If c is a singular 1-cube in $\mathbb{R}^2 - 0$ with $c(0) = c(1)$, show that there is an integer n such that $c - c_{1,n} = \partial c^2$ for some 2-chain c^2 . *Hint:* First partition $[0,1]$ so that each $c([t_{i-1}, t_i])$ is contained on one side of some line through 0.

THE FUNDAMENTAL THEOREM OF CALCULUS

The fact that $d^2 = 0$ and $\partial^2 = 0$, not to mention the typographical similarity of d and ∂ , suggests some connection between chains and forms. This connection is established by integrating forms over chains. Henceforth only differentiable singular n -cubes will be considered.

If ω is a k -form on $[0,1]^k$, then $\omega = f dx^1 \wedge \cdots \wedge dx^k$ for a unique function f . We define

$$\int_{[0,1]^k} \omega = \int_{[0,1]^k} f.$$

We could also write this as

$$\int_{[0,1]^k} f dx^1 \wedge \cdots \wedge dx^k = \int_{[0,1]^k} f(x^1, \dots, x^k) dx^1 \cdots dx^k,$$

one of the reasons for introducing the functions x^i .

If ω is a k -form on A and c is a singular k -cube in A , we define

$$\int_c \omega = \int_{[0,1]^k} c^* \omega.$$

Note, in particular, that

$$\begin{aligned} \int_{I^k} f dx^1 \wedge \cdots \wedge dx^k &= \int_{[0,1]^k} (I^k)^*(f dx^1 \wedge \cdots \wedge dx^k) \\ &= \int_{[0,1]^k} f(x^1, \dots, x^k) dx^1 \cdots dx^k. \end{aligned}$$

A special definition must be made for $k = 0$. A 0-form ω is a function; if $c: \{0\} \rightarrow A$ is a singular 0-cube in A we define

$$\int_c \omega = \omega(c(0)).$$

The integral of ω over a k -chain $c = \sum a_i c_i$ is defined by

$$\int_c \omega = \sum a_i \int_{c_i} \omega.$$

The integral of a 1-form over a 1-chain is often called a **line integral**. If $P dx + Q dy$ is a 1-form on \mathbf{R}^2 and $c: [0,1] \rightarrow \mathbf{R}^2$ is a singular 1-cube (a curve), then one can (but we will not) prove that

$$\begin{aligned} \int_c P dx + Q dy &= \lim \sum_{i=1}^n [c^1(t_i) - c^1(t_{i-1})] \cdot P(c(t_i)) \\ &\quad + [c^2(t_i) - c^2(t_{i-1})] \cdot Q(c(t_i)) \end{aligned}$$

where t_0, \dots, t_n is a partition of $[0,1]$, the choice of t^i in $[t_{i-1}, t_i]$ is arbitrary, and the limit is taken over all partitions

as the maximum of $|t_i - t_{i-1}|$ goes to 0. The right side is often taken as a definition of $\int_c P dx + Q dy$. This is a natural definition to make, since these sums are very much like the sums appearing in the definition of ordinary integrals. However such an expression is almost impossible to work with and is quickly equated with an integral equivalent to $\int_{[0,1]} c^*(P dx + Q dy)$. Analogous definitions for **surface integrals**, that is, integrals of 2-forms over singular 2-cubes, are even more complicated and difficult to use. This is one reason why we have avoided such an approach. The other reason is that the definition given here is the one that makes sense in the more general situations considered in Chapter 5.

The relationship between forms, chains, d , and ∂ is summed up in the neatest possible way by Stokes' theorem, sometimes called the fundamental theorem of calculus in higher dimensions (if $k = 1$ and $c = I^1$, it really is the fundamental theorem of calculus).

4-13 Theorem (Stokes' Theorem). *If ω is a $(k - 1)$ -form on an open set $A \subset \mathbf{R}^n$ and c is a k -chain in A , then*

$$\int_c d\omega = \int_{\partial c} \omega.$$

Proof. Suppose first that $c = I^k$ and ω is a $(k - 1)$ -form on $[0, 1]^k$. Then ω is the sum of $(k - 1)$ -forms of the type

$$f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k,$$

and it suffices to prove the theorem for each of these. This simply involves a computation:

Note that

$$\begin{aligned} & \int_{[0,1]^{k-1}} I_{(j,\alpha)}^k (f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\ &= \begin{cases} 0 & \text{if } j \neq i, \\ \int_{[0,1]^k} f(x^1, \dots, \alpha, \dots, x^k) dx^1 \cdots dx^k & \text{if } j = i. \end{cases} \end{aligned}$$

Therefore

$$\begin{aligned}
& \int_{\partial I^k} f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k \\
&= \sum_{j=1}^k \sum_{\alpha=0,1} (-1)^{j+\alpha} \int_{[0,1]^{k-1}} I_{(j,\alpha)}^k * (f dx^1 \wedge \cdots \wedge \widehat{dx^i} \\
&\quad \wedge \cdots \wedge dx^k) \\
&= (-1)^{i+1} \int_{[0,1]^k} f(x^1, \dots, 1, \dots, x^k) dx^1 \cdots dx^k \\
&\quad + (-1)^i \int_{[0,1]^k} f(x^1, \dots, 0, \dots, x^k) dx^1 \cdots dx^k.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
& \int_{I^k} d(f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\
&= \int_{[0,1]^k} D_i f dx^i \wedge dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k \\
&= (-1)^{i-1} \int_{[0,1]^k} D_i f.
\end{aligned}$$

By Fubini's theorem and the fundamental theorem of calculus (in one dimension) we have

$$\begin{aligned}
& \int_{I^k} d(f dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^k) \\
&= (-1)^{i-1} \int_0^1 \cdots \left(\int_0^1 D_i f(x^1, \dots, x^k) dx^i \right) dx^1 \cdots \\
&\quad \widehat{dx^i} \cdots dx^k \\
&= (-1)^{i-1} \int_0^1 \cdots \int_0^1 [f(x^1, \dots, 1, \dots, x^k) \\
&\quad - f(x^1, \dots, 0, \dots, x^k)] dx^1 \cdots \widehat{dx^i} \cdots dx^k \\
&= (-1)^{i-1} \int_{[0,1]^k} f(x^1, \dots, 1, \dots, x^k) dx^1 \cdots dx^k \\
&\quad + (-1)^i \int_{[0,1]^k} f(x^1, \dots, 0, \dots, x^k) dx^1 \cdots dx^k.
\end{aligned}$$

Thus

$$\int_{I^k} d\omega = \int_{\partial I^k} \omega.$$

If c is an arbitrary singular k -cube, working through the definitions will show that

$$\int_{\partial c} \omega = \int_{\partial I^k} c^* \omega.$$

Therefore

$$\int_c d\omega = \int_{I^k} c^*(d\omega) = \int_{I^k} d(c^*\omega) = \int_{\partial I^k} c^*\omega = \int_{\partial c} \omega.$$

Finally, if c is a k -chain $\sum a_i c_i$, we have

$$\int_c d\omega = \sum a_i \int_{c_i} d\omega = \sum a_i \int_{\partial c_i} \omega = \int_{\partial c} \omega. \blacksquare$$

Stokes' theorem shares three important attributes with many fully evolved major theorems:

1. It is trivial.
2. It is trivial because the terms appearing in it have been properly defined.
3. It has significant consequences.

Since this entire chapter was little more than a series of definitions which made the statement and proof of Stokes' theorem possible, the reader should be willing to grant the first two of these attributes to Stokes' theorem. The rest of the book is devoted to justifying the third.

Problems. 4-25. (*Independence of parameterization*). Let c be a singular k -cube and $p: [0,1]^k \rightarrow [0,1]^k$ a 1-1 function such that $p([0,1]^k) = [0,1]^k$ and $\det p'(x) \geq 0$ for $x \in [0,1]^k$. If ω is a k -form, show that

$$\int_c \omega = \int_{c \circ p} \omega.$$

- 4-26. Show that $\int_{c_{R,n}} d\theta = 2\pi n$, and use Stokes' theorem to conclude that $c_{R,n} \neq \partial c$ for any 2-chain c in $\mathbf{R}^2 - 0$ (recall the definition of $c_{R,n}$ in Problem 4-23).
- 4-27. Show that the integer n of Problem 4-24 is unique. This integer is called the **winding number** of c around 0.
- 4-28. Recall that the set of complex numbers \mathbf{C} is simply \mathbf{R}^2 with $(a,b) = a + bi$. If $a_1, \dots, a_n \in \mathbf{C}$ let $f: \mathbf{C} \rightarrow \mathbf{C}$ be $f(z) = z^n + a_1 z^{n-1} + \dots + a_n$. Define the singular 1-cube $c_{R,f}$:

$[0,1] \rightarrow \mathbf{C} - 0$ by $c_{R,f} = f \circ c_{R,1}$, and the singular 2-cube c by $c(s,t) = t \cdot c_{R,n}(s) + (1-t)c_{R,f}(s)$.

(a) Show that $\partial c = c_{R,f} - c_{R,n}$, and that $c([0,1] \times [0,1]) \subset \mathbf{C} - 0$ if R is large enough.

(b) Using Problem 4-26, prove the *Fundamental Theorem of Algebra*: Every polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$ with $a_i \in \mathbf{C}$ has a root in \mathbf{C} .

4-29. If ω is a 1-form $f dx$ on $[0,1]$ with $f(0) = f(1)$, show that there is a unique number λ such that $\omega - \lambda dx = dg$ for some function g with $g(0) = g(1)$. *Hint*: Integrate $\omega - \lambda dx = dg$ on $[0,1]$ to find λ .

4-30. If ω is a 1-form on $\mathbf{R}^2 - 0$ such that $d\omega = 0$, prove that

$$\omega = \lambda d\theta + dg$$

for some $\lambda \in \mathbf{R}$ and $g: \mathbf{R}^2 - 0 \rightarrow \mathbf{R}$. *Hint*: If

$$c_{R,1}^*(\omega) = \lambda_R dx + d(g_R),$$

show that all numbers λ_R have the same value λ .

4-31. If $\omega \neq 0$, show that there is a chain c such that $\int_c \omega \neq 0$. Use this fact, Stokes' theorem and $\partial^2 = 0$ to prove $d^2 = 0$.

4-32. (a) Let c_1, c_2 be singular 1-cubes in \mathbf{R}^2 with $c_1(0) = c_2(0)$ and $c_1(1) = c_2(1)$. Show that there is a singular 2-cube c such that $\partial c = c_1 - c_2 + c_3 - c_4$, where c_3 and c_4 are *degenerate*, that is, $c_3([0,1])$ and $c_4([0,1])$ are points. Conclude that $\int_{c_1} \omega = \int_{c_2} \omega$ if ω is exact. Give a counterexample on $\mathbf{R}^2 - 0$ if ω is merely closed.

(b) If ω is a 1-form on a subset of \mathbf{R}^2 and $\int_{c_1} \omega = \int_{c_2} \omega$ for all c_1, c_2 with $c_1(0) = c_2(0)$ and $c_1(1) = c_2(1)$, show that ω is exact. *Hint*: Consider Problems 2-21 and 3-34.

4-33. (*A first course in complex variables.*) If $f: \mathbf{C} \rightarrow \mathbf{C}$, define f to be **differentiable** at $z_0 \in \mathbf{C}$ if the limit

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. (This quotient involves two complex numbers and this definition is completely different from the one in Chapter 2.) If f is differentiable at every point z in an open set A and f' is continuous on A , then f is called **analytic** on A .

(a) Show that $f(z) = z$ is analytic and $f(z) = \bar{z}$ is not (where $\overline{x + iy} = x - iy$). Show that the sum, product, and quotient of analytic functions are analytic.

(b) If $f = u + iv$ is analytic on A , show that u and v satisfy the *Cauchy-Riemann equations*:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

Hint: Use the fact that $\lim_{z \rightarrow z_0} [f(z) - f(z_0)]/(z - z_0)$ must be the same for $z = z_0 + (x + i \cdot 0)$ and $z = z_0 + (0 + i \cdot y)$ with $x, y \rightarrow 0$. (The converse is also true, if u and v are continuously differentiable; this is more difficult to prove.)

(c) Let $T: \mathbf{C} \rightarrow \mathbf{C}$ be a linear transformation (where \mathbf{C} is considered as a vector space over \mathbf{R}). If the matrix of T with respect to the basis $(1, i)$ is $\begin{pmatrix} a, b \\ c, d \end{pmatrix}$ show that T is multiplication by a complex number if and only if $a = d$ and $b = -c$. Part (b) shows that an analytic function $f: \mathbf{C} \rightarrow \mathbf{C}$, considered as a function $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$, has a derivative $Df(z_0)$ which is multiplication by a complex number. What complex number is this?

(d) Define

$$\begin{aligned} d(\omega + i\eta) &= d\omega + i d\eta, \\ \int_c \omega + i\eta &= \int_c \omega + i \int_c \eta, \end{aligned}$$

$$(\omega + i\eta) \wedge (\theta + i\lambda) = \omega \wedge \theta - \eta \wedge \lambda + i(\eta \wedge \theta + \omega \wedge \lambda),$$

and

$$dz = dx + i dy.$$

Show that $d(f \cdot dz) = 0$ if and only if f satisfies the Cauchy-Riemann equations.

(e) Prove the *Cauchy Integral Theorem*: If f is analytic on A , then $\int_c f dz = 0$ for every closed curve c (singular 1-cube with $c(0) = c(1)$) such that $c = \partial c'$ for some 2-chain c' in A .

(f) Show that if $g(z) = 1/z$, then $g \cdot dz$ [or $(1/z)dz$ in classical notation] equals $i d\theta + dh$ for some function $h: \mathbf{C} - 0 \rightarrow \mathbf{R}$. Conclude that $\int_{c_{R,n}} (1/z) dz = 2\pi i n$.

(g) If f is analytic on $\{z: |z| < 1\}$, use the fact that $g(z) = f(z)/z$ is analytic in $\{z: 0 < |z| < 1\}$ to show that

$$\int_{c_{R_1,n}} \frac{f(z)}{z} dz = \int_{c_{R_2,n}} \frac{f(z)}{z} dz$$

if $0 < R_1, R_2 < 1$. Use (f) to evaluate $\lim_{R \rightarrow 0} \int_{c_{R,n}} f(z)/z dz$ and conclude:

Cauchy Integral Formula: If f is analytic on $\{z: |z| < 1\}$ and c is a closed curve in $\{z: 0 < |z| < 1\}$ with winding number n around 0, then

$$n \cdot f(0) = \frac{1}{2\pi i} \int_c \frac{f(z)}{z} dz.$$

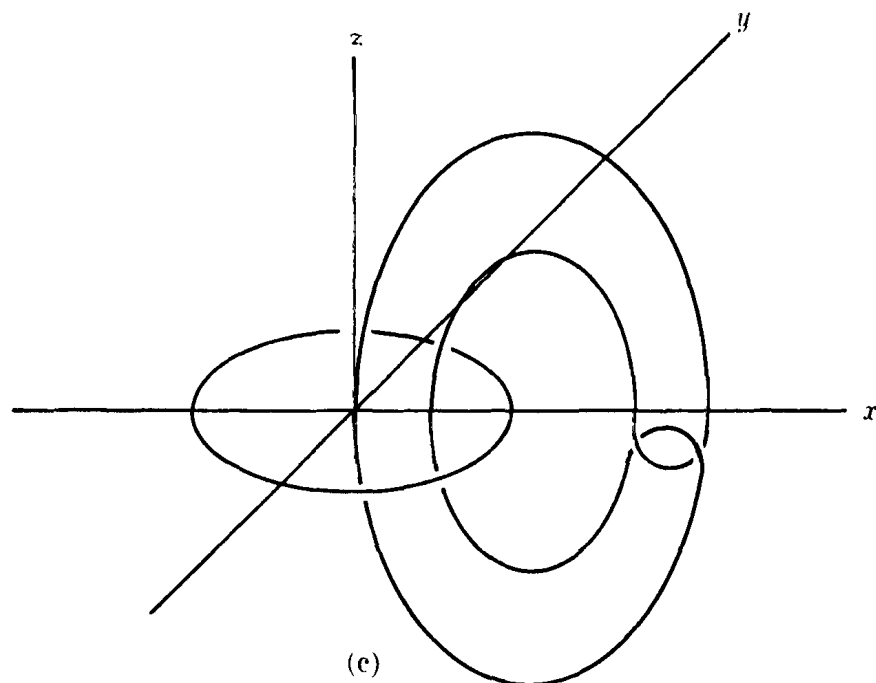
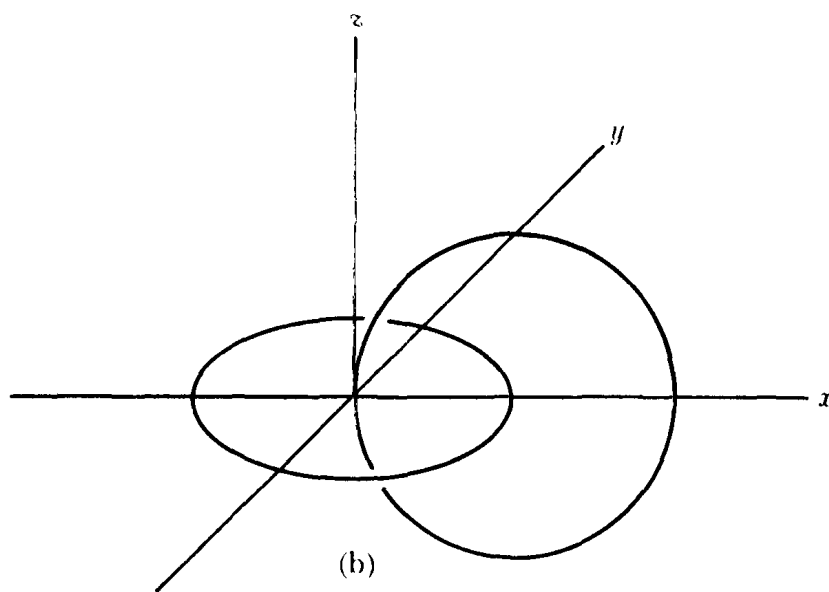
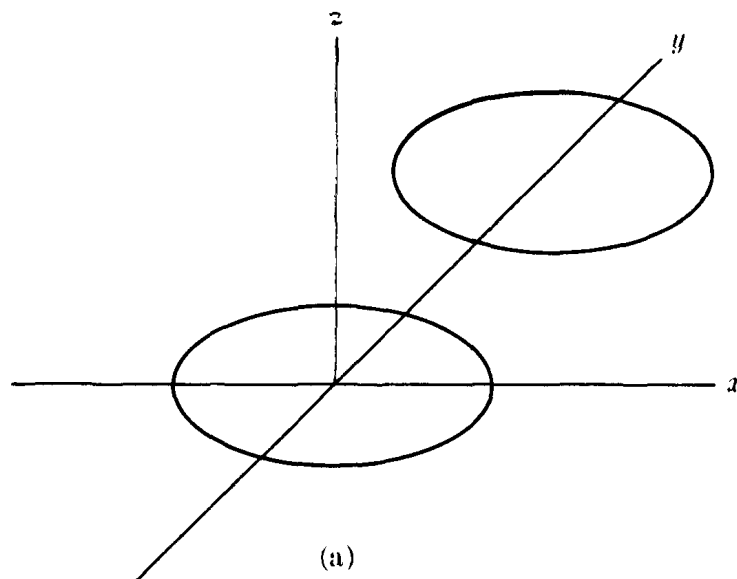


FIGURE 4-6

4-34. If $F: [0, 1]^2 \rightarrow \mathbf{R}^3$ and $s \in [0, 1]$ define $F_s: [0, 1] \rightarrow \mathbf{R}^3$ by $F_s(t) = F(s, t)$. If each F_s is a closed curve, F is called a **homotopy** between the closed curve F_0 and the closed curve F_1 . Suppose F and G are homotopies of closed curves; if for each s the closed curves F_s and G_s do not intersect, the pair (F, G) is called a homotopy between the nonintersecting closed curves F_0, G_0 and F_1, G_1 . It is intuitively obvious that there is no such homotopy with F_0, G_0 the pair of curves shown in Figure 4-6 (a), and F_1, G_1 the pair of (b) or (c). The present problem, and Problem 5-33 prove this for (b) but the proof for (c) requires different techniques.

(a) If $f, g: [0, 1] \rightarrow \mathbf{R}^3$ are nonintersecting closed curves define $c_{f,g}: [0, 1]^2 \rightarrow \mathbf{R}^3 - 0$ by

$$c_{f,g}(u, v) = f(u) - g(v).$$

If (F, G) is a homotopy of nonintersecting closed curves define $C_{F,G}: [0, 1]^3 \rightarrow \mathbf{R}^3 - 0$ by

$$C_{F,G}(s, u, v) = c_{F_s, G_s}(u, v) = F(s, u) - G(s, v).$$

Show that

$$\partial C_{F,G} = c_{F_0, G_0} - c_{F_1, G_1}.$$

(b) If ω is a closed 2-form on $\mathbf{R}^3 - 0$ show that

$$\int_{c_{F_0, G_0}} \omega = \int_{c_{F_1, G_1}} \omega.$$

5

Integration on Manifolds

MANIFOLDS

If U and V are open sets in \mathbf{R}^n , a differentiable function $h: U \rightarrow V$ with a differentiable inverse $h^{-1}: V \rightarrow U$ will be called a **diffeomorphism**. (“Differentiable” henceforth means “ C^∞ ”.)

A subset M of \mathbf{R}^n is called a **k -dimensional manifold** (in \mathbf{R}^n) if for every point $x \in M$ the following condition is satisfied:

(M) There is an open set U containing x , an open set $V \subset \mathbf{R}^n$, and a diffeomorphism $h: U \rightarrow V$ such that

$$\begin{aligned} h(U \cap M) &= V \cap (\mathbf{R}^k \times \{0\}) \\ &= \{y \in V: y^{k+1} = \cdots = y^n = 0\}. \end{aligned}$$

In other words, $U \cap M$ is, “up to diffeomorphism,” simply $\mathbf{R}^k \times \{0\}$ (see Figure 5-1). The two extreme cases of our definition should be noted: a point in \mathbf{R}^n is a 0-dimensional manifold, and an open subset of \mathbf{R}^n is an n -dimensional manifold.

One common example of an n -dimensional manifold is the

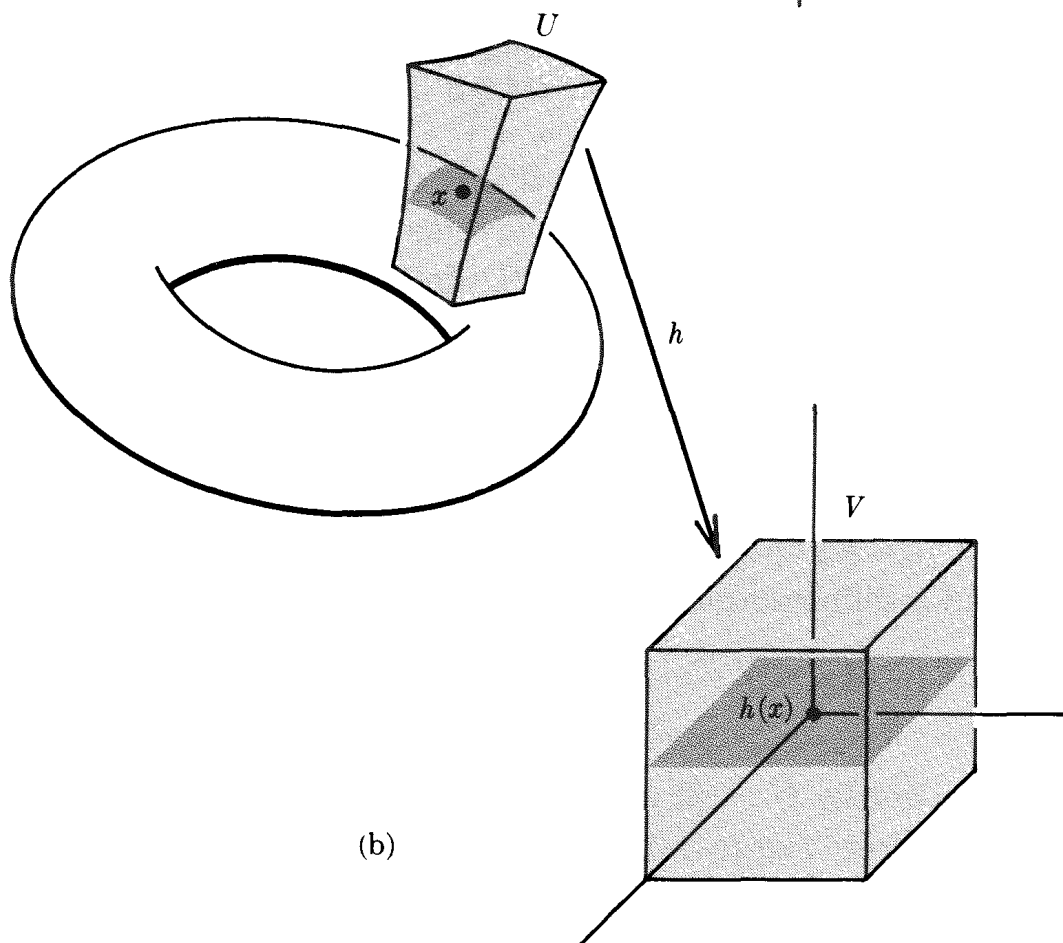
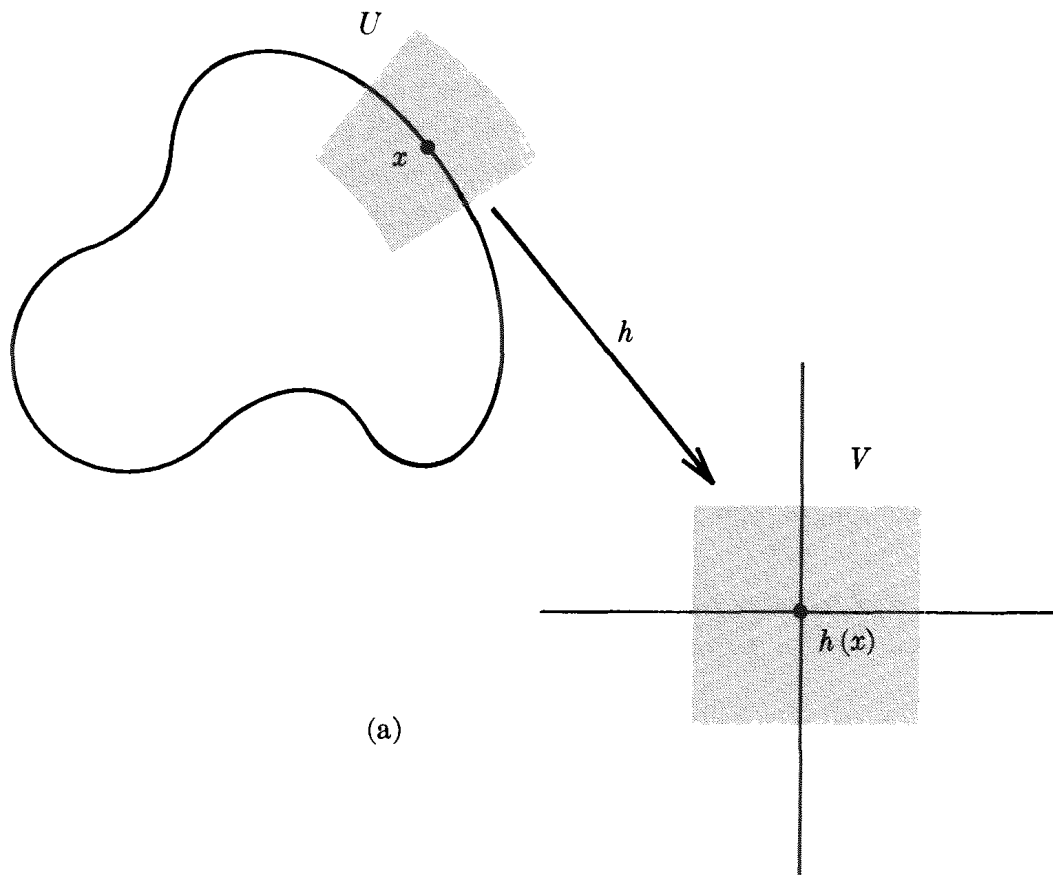


FIGURE 5-1. A one-dimensional manifold in \mathbb{R}^2 and a two-dimensional manifold in \mathbb{R}^3 .

n -sphere S^n , defined as $\{x \in \mathbf{R}^{n+1}: |x| = 1\}$. We leave it as an exercise for the reader to prove that condition (M) is satisfied. If you are unwilling to trouble yourself with the details, you may instead use the following theorem, which provides many examples of manifolds (note that $S^n = g^{-1}(0)$, where $g: \mathbf{R}^{n+1} \rightarrow \mathbf{R}$ is defined by $g(x) = |x|^2 - 1$).

5-1 Theorem. *Let $A \subset \mathbf{R}^n$ be open and let $g: A \rightarrow \mathbf{R}^p$ be a differentiable function such that $g'(x)$ has rank p whenever $g(x) = 0$. Then $g^{-1}(0)$ is an $(n - p)$ -dimensional manifold in \mathbf{R}^n .*

Proof. This follows immediately from Theorem 2-13. ■

There is an alternative characterization of manifolds which is very important.

5-2 Theorem. *A subset M of \mathbf{R}^n is a k -dimensional manifold if and only if for each point $x \in M$ the following “coordinate condition” is satisfied:*

(C) *There is an open set U containing x , an open set $W \subset \mathbf{R}^k$, and a 1-1 differentiable function $f: W \rightarrow \mathbf{R}^n$ such that*

- (1) $f(W) = M \cap U$,
- (2) $f'(y)$ has rank k for each $y \in W$,
- (3) $f^{-1}: f(W) \rightarrow W$ is continuous.

[Such a function f is called a **coordinate system** around x (see Figure 5-2).]

Proof. If M is a k -dimensional manifold in \mathbf{R}^n , choose $h: U \rightarrow V$ satisfying (M). Let $W = \{a \in \mathbf{R}^k: (a, 0) \in h(M)\}$ and define $f: W \rightarrow \mathbf{R}^n$ by $f(a) = h^{-1}(a, 0)$. Clearly $f(W) = M \cap U$ and f^{-1} is continuous. If $H: U \rightarrow \mathbf{R}^k$ is $H(z) = (h^1(z), \dots, h^k(z))$, then $H(f(y)) = y$ for all $y \in W$; therefore $H'(f(y)) \cdot f'(y) = I$ and $f'(y)$ must have rank k .

Suppose, conversely, that $f: W \rightarrow \mathbf{R}^n$ satisfies condition (C). Let $x = f(y)$. Assume that the matrix $(D_j f^i(y))$, $1 \leq i, j \leq k$ has a non-zero determinant. Define $g: W \times \mathbf{R}^{n-k} \rightarrow \mathbf{R}^n$ by

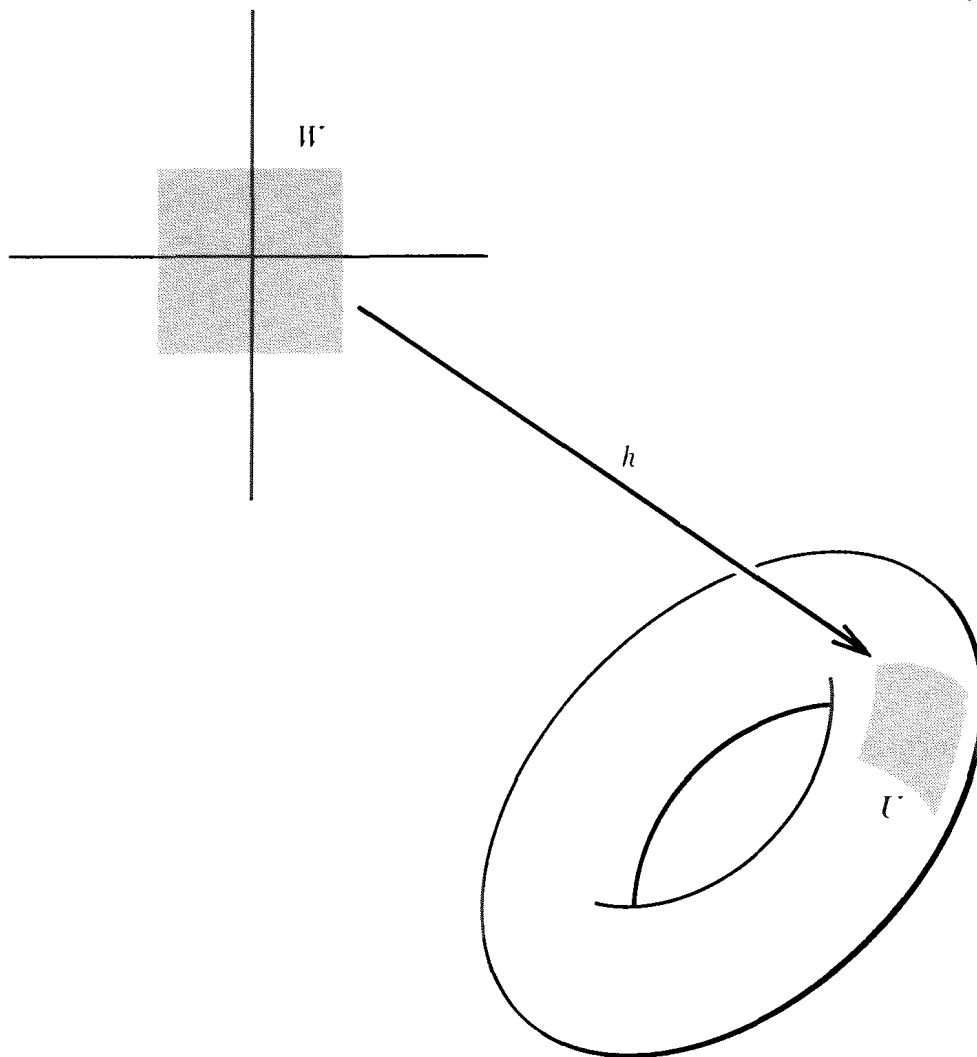


FIGURE 5-2

$g(a,b) = f(a) + (0,b)$. Then $\det g'(a,b) = \det (D_j f^i(a))$, so $\det g'(y,0) \neq 0$. By Theorem 2-11 there is an open set V_1' containing $(y,0)$ and an open set V_2' containing $g(y,0) = x$ such that $g: V_1' \rightarrow V_2'$ has a differentiable inverse $h: V_2' \rightarrow V_1'$. Since f^{-1} is continuous, $\{f(a): (a,0) \in V_1'\} = U \cap f(W)$ for some open set U . Let $V_2 = V_2' \cap U$ and $V_1 = g^{-1}(V_2)$. Then $V_2 \cap M$ is exactly $\{f(a): (a,0) \in V_1\} = \{g(a,0): (a,0) \in V_1\}$, so

$$\begin{aligned} h(V_2 \cap M) &= g^{-1}(V_2 \cap M) = g^{-1}(\{g(a,0): (a,0) \in V_1\}) \\ &= V_1 \cap (\mathbf{R}^k \times \{0\}). \quad \blacksquare \end{aligned}$$

One consequence of the proof of Theorem 5-2 should be noted. If $f_1: W_1 \rightarrow \mathbf{R}^n$ and $f_2: W_2 \rightarrow \mathbf{R}^n$ are two coordinate

systems, then

$$f_2^{-1} \circ f_1: f_1^{-1}(f_2(W_2)) \rightarrow \mathbf{R}^k$$

is differentiable with non-singular Jacobian. In fact, $f_2^{-1}(y)$ consists of the first k components of $h(y)$.

The **half-space** $\mathbf{H}^k \subset \mathbf{R}^k$ is defined as $\{x \in \mathbf{R}^k: x^k \geq 0\}$. A subset M of \mathbf{R}^n is a **k -dimensional manifold-with-boundary** (Figure 5-3) if for every point $x \in M$ either condition (M) or the following condition is satisfied:

(M') There is an open set U containing x , an open set $V \subset \mathbf{R}^n$, and a diffeomorphism $h: U \rightarrow V$ such that

$$\begin{aligned} h(U \cap M) &= V \cap (\mathbf{H}^k \times \{0\}) \\ &= \{y \in V: y^k \geq 0 \text{ and } y^{k+1} = \dots = y^n = 0\} \end{aligned}$$

and $h(x)$ has k th component $= 0$.

It is important to note that conditions (M) and (M') cannot both hold for the same x . In fact, if $h_1: U_1 \rightarrow V_1$ and $h_2: U_2 \rightarrow V_2$ satisfied (M) and (M'), respectively, then $h_2 \circ h_1^{-1}$ would be a differentiable map that takes an open set in \mathbf{R}^k , containing $h(x)$, into a subset of \mathbf{H}^k which is not open in \mathbf{R}^k . Since $\det(h_2 \circ h_1^{-1})' \neq 0$, this contradicts Problem 2-36. The set of all points $x \in M$ for which condition M' is satisfied is called the **boundary** of M and denoted ∂M . This

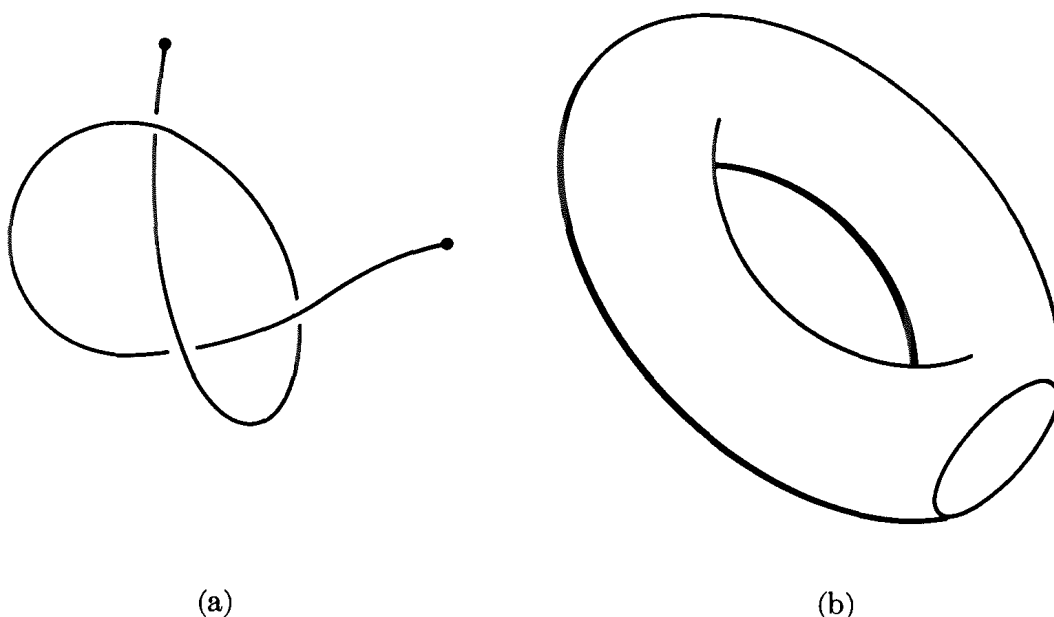


FIGURE 5-3. A one-dimensional and a two-dimensional manifold-with-boundary in \mathbf{R}^3 .

must not be confused with the boundary of a set, as defined in Chapter 1 (see Problems 5-3 and 5-8).

Problems. 5-1. If M is a k -dimensional manifold-with-boundary, prove that ∂M is a $(k - 1)$ -dimensional manifold and $M - \partial M$ is a k -dimensional manifold.

5-2. Find a counterexample to Theorem 5-2 if condition (3) is omitted.

Hint: Wrap an open interval into a figure six.

5-3. (a) Let $A \subset \mathbf{R}^n$ be an open set such that boundary A is an $(n - 1)$ -dimensional manifold. Show that $N = A \cup \text{boundary } A$ is an n -dimensional manifold-with-boundary. (It is well to bear in mind the following example: if $A = \{x \in \mathbf{R}^n: |x| < 1 \text{ or } 1 < |x| < 2\}$ then $N = A \cup \text{boundary } A$ is a manifold-with-boundary, but $\partial N \neq \text{boundary } A$.)

(b) Prove a similar assertion for an open subset of an n -dimensional manifold.

5-4. Prove a partial converse of Theorem 5-1: If $M \subset \mathbf{R}^n$ is a k -dimensional manifold and $x \in M$, then there is an open set $A \subset \mathbf{R}^n$ containing x and a differentiable function $g: A \rightarrow \mathbf{R}^{n-k}$ such that $A \cap M = g^{-1}(0)$ and $g'(y)$ has rank $n - k$ when $g(y) = 0$.

5-5. Prove that a k -dimensional (vector) subspace of \mathbf{R}^n is a k -dimensional manifold.

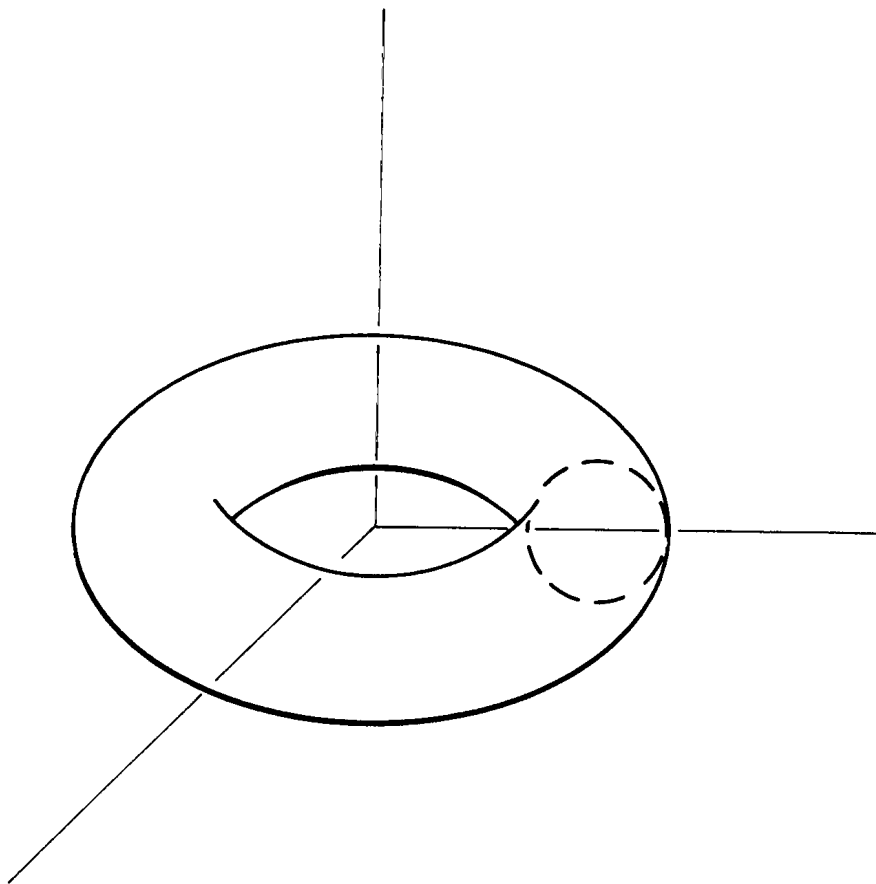


FIGURE 5-4

- 5-6. If $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$, the **graph** of f is $\{(x, y): y = f(x)\}$. Show that the graph of f is an n -dimensional manifold if and only if f is differentiable.
- 5-7. Let $\mathbf{K}^n = \{x \in \mathbf{R}^n: x^1 = 0 \text{ and } x^2, \dots, x^{n-1} > 0\}$. If $M \subset \mathbf{K}^n$ is a k -dimensional manifold and N is obtained by revolving M around the axis $x^1 = \dots = x^{n-1} = 0$, show that N is a $(k + 1)$ -dimensional manifold. **Example: the torus (Figure 5-4).**
- 5-8. (a) If M is a k -dimensional manifold in \mathbf{R}^n and $k < n$, show that M has measure 0.
- (b) If M is a closed n -dimensional manifold-with-boundary in \mathbf{R}^n , show that the boundary of M is ∂M . Give a counterexample if M is not closed.
- (c) If M is a compact n -dimensional manifold-with-boundary in \mathbf{R}^n , show that M is Jordan-measurable.

FIELDS AND FORMS ON MANIFOLDS

Let M be a k -dimensional manifold in \mathbf{R}^n and let $f: W \rightarrow \mathbf{R}^n$ be a coordinate system around $x = f(a)$. Since $f'(a)$ has rank k , the linear transformation $f_*: \mathbf{R}^k_a \rightarrow \mathbf{R}^n_x$ is 1-1, and $f_*(\mathbf{R}^k_a)$ is a k -dimensional subspace of \mathbf{R}^n_x . If $g: V \rightarrow \mathbf{R}^n$ is another coordinate system, with $x = g(b)$, then

$$g_*(\mathbf{R}^k_b) = f_*(f^{-1} \circ g)_*(\mathbf{R}^k_b) = f_*(\mathbf{R}^k_a).$$

Thus the k -dimensional subspace $f_*(\mathbf{R}^k_a)$ does not depend on the coordinate system f . This subspace is denoted M_x , and is called the **tangent space** of M at x (see Figure 5-5). In later sections we will use the fact that there is a natural inner product T_x on M_x , induced by that on \mathbf{R}^n_x : if $v, w \in M_x$ define $T_x(v, w) = \langle v, w \rangle_x$.

Suppose that A is an open set containing M , and F is a differentiable vector field on A such that $F(x) \in M_x$ for each $x \in M$. If $f: W \rightarrow \mathbf{R}^n$ is a coordinate system, there is a unique (differentiable) vector field G on W such that $f_*(G(a)) = F(f(a))$ for each $a \in W$. We can also consider a function F which merely assigns a vector $F(x) \in M_x$ for each $x \in M$; such a function is called a **vector field on M** . There is still a unique vector field G on W such that $f_*(G(a)) = F(f(a))$ for $a \in W$; we *define* F to be differentiable if G is differentiable. Note that our definition does not depend on the coordinate

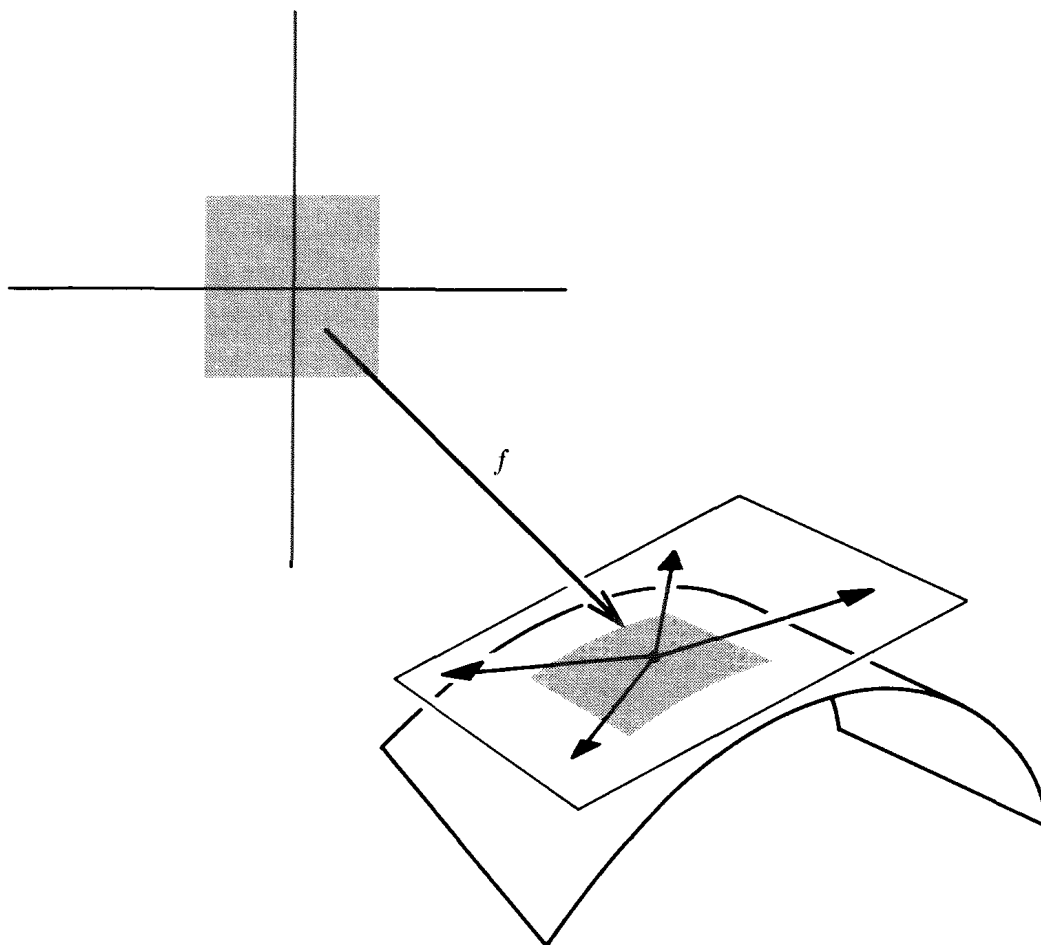


FIGURE 5-5

system chosen: if $g: V \rightarrow \mathbf{R}^n$ and $g_*(H(b)) = F(g(b))$ for all $b \in V$, then the component functions of $H(b)$ must equal the component functions of $G(f^{-1}(g(b)))$, so H is differentiable if G is.

Precisely the same considerations hold for forms. A function ω which assigns $\omega(x) \in \Lambda^p(M_x)$ for each $x \in M$ is called a **p -form on M** . If $f: W \rightarrow \mathbf{R}^n$ is a coordinate system, then $f^*\omega$ is a p -form on W ; we *define* ω to be differentiable if $f^*\omega$ is. A p -form ω on M can be written as

$$\omega = \sum_{i_1 < \cdots < i_p} \omega_{i_1, \dots, i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p}.$$

Here the functions ω_{i_1, \dots, i_p} are defined only on M . The definition of $d\omega$ given previously would make no sense here, since $D_j(\omega_{i_1, \dots, i_p})$ has no meaning. Nevertheless, there is a reasonable way of defining $d\omega$.

5-3 Theorem. *There is a unique $(p + 1)$ -form $d\omega$ on M such that for every coordinate system $f: W \rightarrow \mathbf{R}^n$ we have*

$$f^*(d\omega) = d(f^*\omega).$$

Proof. If $f: W \rightarrow \mathbf{R}^n$ is a coordinate system with $x = f(a)$ and $v_1, \dots, v_{p+1} \in M_x$, there are unique w_1, \dots, w_{p+1} in \mathbf{R}^k_a such that $f_*(w_i) = v_i$. Define $d\omega(x)(v_1, \dots, v_{p+1}) = d(f^*\omega)(a)(w_1, \dots, w_{p+1})$. One can check that this definition of $d\omega(x)$ does not depend on the coordinate system f , so that $d\omega$ is well-defined. Moreover, it is clear that $d\omega$ has to be defined this way, so $d\omega$ is unique. ■

It is often necessary to choose an orientation μ_x for each tangent space M_x of a manifold M . Such choices are called **consistent** (Figure 5-6) provided that for every coordinate

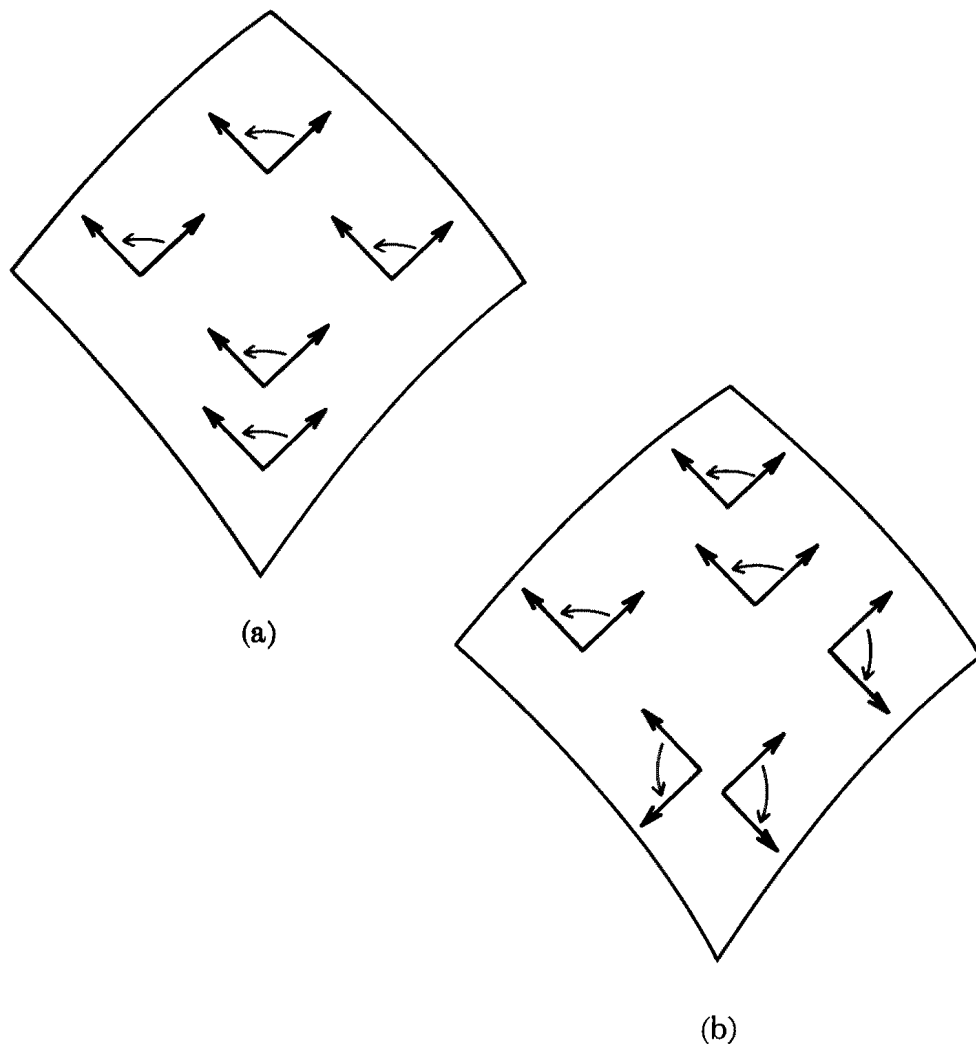


FIGURE 5-6. (a) Consistent and (b) inconsistent choices of orientations.

system $f: W \rightarrow \mathbf{R}^n$ and $a, b \in W$ the relation

$$[f_*((e_1)_a), \dots, f_*((e_k)_a)] = \mu_{f(a)}$$

holds if and only if

$$[f_*((e_1)_b), \dots, f_*((e_k)_b)] = \mu_{f(b)}.$$

Suppose orientations μ_x have been chosen consistently. If $f: W \rightarrow \mathbf{R}^n$ is a coordinate system such that

$$[f_*((e_1)_a), \dots, f_*((e_k)_a)] = \mu_{f(a)}$$

for one, and hence for every $a \in W$, then f is called **orientation-preserving**. If f is *not* orientation-preserving and $T: \mathbf{R}^k \rightarrow \mathbf{R}^k$ is a linear transformation with $\det T = -1$, then $f \circ T$ is orientation-preserving. Therefore there is an orientation-preserving coordinate system around each point. If f and g are orientation-preserving and $x = f(a) = g(b)$, then the relation

$$[f_*((e_1)_a), \dots, f_*((e_k)_a)] = \mu_x = [g_*((e_1)_b), \dots, g_*((e_k)_b)]$$

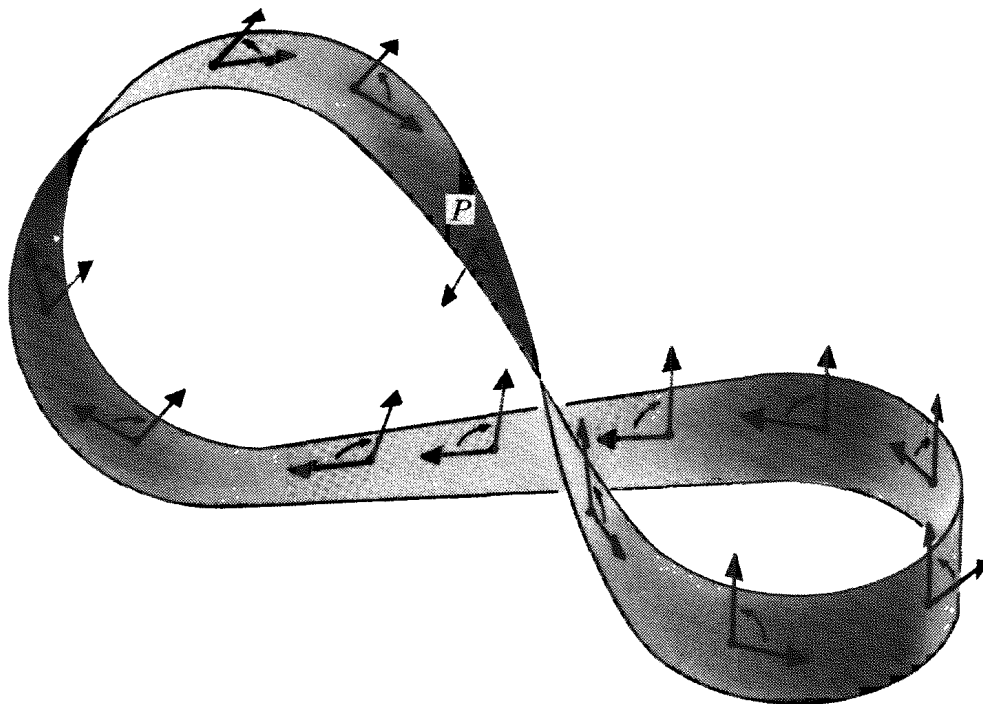


FIGURE 5-7. The Möbius strip, a non-orientable manifold. A basis begins at P , moves to the right and around, and comes back to P with the wrong orientation.

implies that

$$[(g^{-1} \circ f)_*((e_1)_a), \dots, (g^{-1} \circ f)_*((e_k)_a)] = [(e_1)_b, \dots, (e_k)_b],$$

so that $\det (g^{-1} \circ f)' > 0$, an important fact to remember.

A manifold for which orientations μ_x can be chosen consistently is called **orientable**, and a particular choice of the μ_x is called an **orientation** μ of M . A manifold together with an orientation μ is called an **oriented** manifold. The classical example of a non-orientable manifold is the Möbius strip. A model can be made by gluing together the ends of a strip of paper which has been given a half twist (Figure 5-7).

Our definitions of vector fields, forms, and orientations can be made for manifolds-with-boundary also. If M is a k -dimensional manifold-with-boundary and $x \in \partial M$, then $(\partial M)_x$ is a $(k - 1)$ -dimensional subspace of the k -dimensional vector space M_x . Thus there are exactly two unit vectors in M_x which are perpendicular to $(\partial M)_x$; they can be distinguished as follows (Figure 5-8). If $f: W \rightarrow \mathbf{R}^n$ is a coordinate system with $W \subset H^k$ and $f(0) = x$, then only one of these unit vectors is $f_*(v_0)$ for some v_0 with $v^k < 0$. This unit vector is called the **outward unit normal** $n(x)$; it is not hard to check that this definition does not depend on the coordinate system f .

Suppose that μ is an orientation of a k -dimensional manifold-with-boundary M . If $x \in \partial M$, choose $v_1, \dots, v_{k-1} \in (\partial M)_x$ so that $[n(x), v_1, \dots, v_{k-1}] = \mu_x$. If it is also true that $[n(x), w_1, \dots, w_{k-1}] = \mu_x$, then both $[v_1, \dots, v_{k-1}]$ and $[w_1, \dots, w_{k-1}]$ are the same orientation for $(\partial M)_x$. This orientation is denoted $(\partial\mu)_x$. It is easy to see that the orientations $(\partial\mu)_x$, for $x \in \partial M$, are consistent on ∂M . Thus if M is orientable, ∂M is also orientable, and an orientation μ for M determines an orientation $\partial\mu$ for ∂M , called the **induced orientation**. If we apply these definitions to \mathbf{H}^k with the usual orientation, we find that the induced orientation on $\mathbf{R}^{k-1} = \{x \in \mathbf{H}^k: x^k = 0\}$ is $(-1)^k$ times the usual orientation. The reason for such a choice will become clear in the next section.

If M is an *oriented* $(n - 1)$ -dimensional manifold in \mathbf{R}^n , a substitute for outward unit normal vectors can be defined,

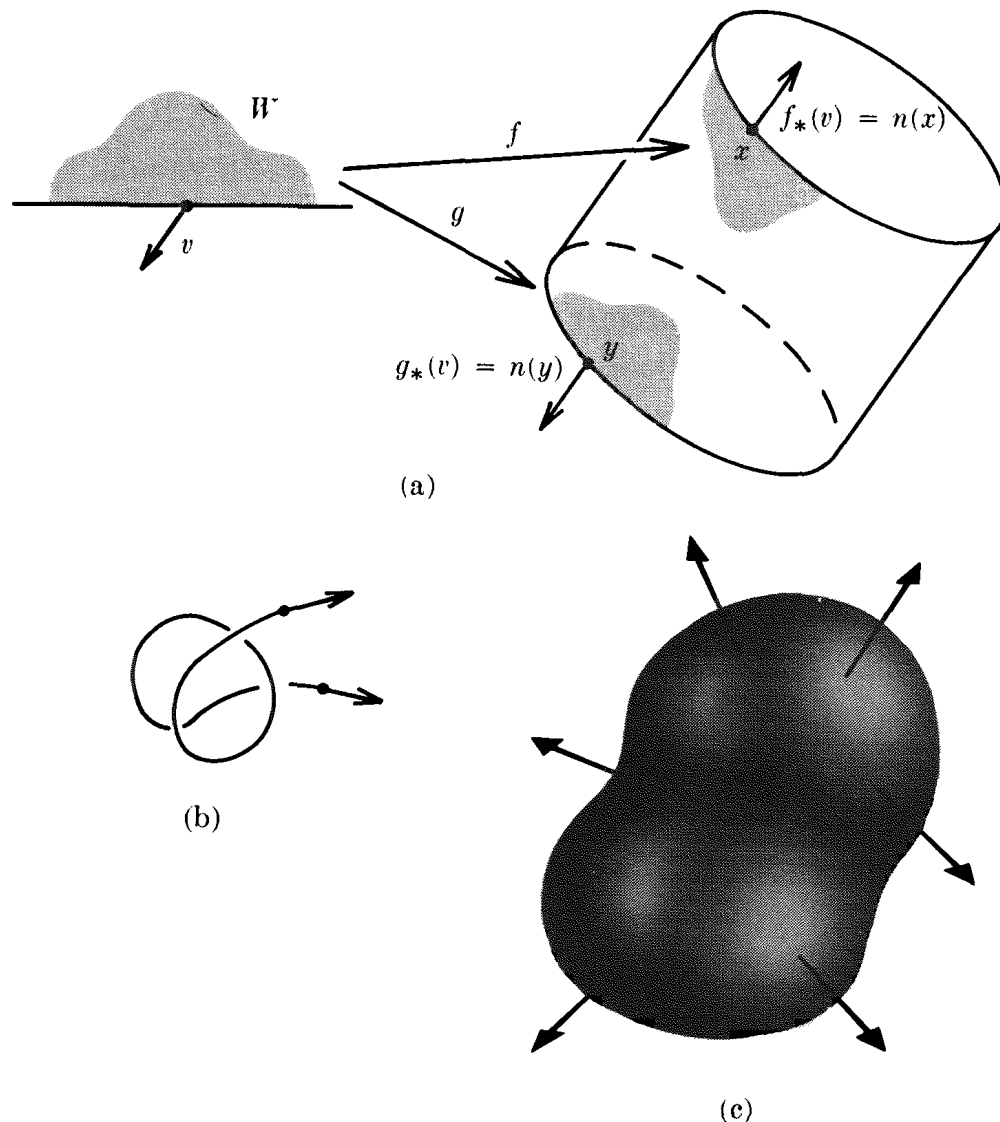


FIGURE 5-8. Some outward unit normal vectors of manifolds-with-boundary in \mathbf{R}^3 .

even though M is not necessarily the boundary of an n -dimensional manifold. If $[v_1, \dots, v_{n-1}] = \mu_x$, we choose $n(x)$ in \mathbf{R}^n_x so that $n(x)$ is a unit vector perpendicular to M_x and $[n(x), v_1, \dots, v_{n-1}]$ is the usual orientation of \mathbf{R}^n_x . We still call $n(x)$ the outward unit normal to M (determined by μ). The vectors $n(x)$ vary continuously on M , in an obvious sense. Conversely, if a continuous family of unit normal vectors $n(x)$ is defined on all of M , then we can determine an orientation of M . This shows that such a continuous choice of normal vectors is impossible on the Möbius strip. In the paper model of the Möbius strip the two sides of the paper (which has thickness) may be thought of as the end points of the unit

normal vectors in both directions. The impossibility of choosing normal vectors continuously is reflected by the famous property of the paper model. The paper model is one-sided (if you start to paint it on one side you end up painting it all over); in other words, choosing $n(x)$ arbitrarily at one point, and then by the continuity requirement at other points, eventually forces the opposite choice for $n(x)$ at the initial point.

Problems. 5-9. Show that M_x consists of the tangent vectors at t of curves c in M with $c(t) = x$.

5-10. Suppose \mathcal{C} is a collection of coordinate systems for M such that (1) For each $x \in M$ there is $f \in \mathcal{C}$ which is a coordinate system around x ; (2) if $f, g \in \mathcal{C}$, then $\det(f^{-1} \circ g)' > 0$. Show that there is a unique orientation of M such that f is orientation-preserving if $f \in \mathcal{C}$.

5-11. If M is an n -dimensional manifold-with-boundary in \mathbf{R}^n , define μ_x as the usual orientation of $M_x = \mathbf{R}^n_x$ (the orientation μ so defined is the **usual orientation** of M). If $x \in \partial M$, show that the two definitions of $n(x)$ given above agree.

5-12. (a) If F is a differentiable vector field on $M \subset \mathbf{R}^n$, show that there is an open set $A \supset M$ and a differentiable vector field \tilde{F} on A with $\tilde{F}(x) = F(x)$ for $x \in M$. *Hint:* Do this locally and use partitions of unity.

(b) If M is closed, show that we can choose $A = \mathbf{R}^n$.

5-13. Let $g: A \rightarrow \mathbf{R}^p$ be as in Theorem 5-1.

(a) If $x \in M = g^{-1}(0)$, let $h: U \rightarrow \mathbf{R}^n$ be the essentially unique diffeomorphism such that $g \circ h(y) = (y^{n-p+1}, \dots, y^n)$ and $h(0) = x$. Define $f: \mathbf{R}^{n-p} \rightarrow \mathbf{R}^n$ by $f(a) = h(0, a)$. Show that f_* is 1-1 so that the $n - p$ vectors $f_*((e_1)_0), \dots, f_*((e_{n-p})_0)$ are linearly independent.

(b) Show that orientations μ_x can be defined consistently, so that M is orientable.

(c) If $p = 1$, show that the components of the outward normal at x are some multiple of $D_1g(x), \dots, D_n g(x)$.

5-14. If $M \subset \mathbf{R}^n$ is an orientable $(n - 1)$ -dimensional manifold, show that there is an open set $A \subset \mathbf{R}^n$ and a differentiable $g: A \rightarrow \mathbf{R}^1$ so that $M = g^{-1}(0)$ and $g'(x)$ has rank 1 for $x \in M$. *Hint:* Problem 5-4 does this locally. Use the orientation to choose consistent local solutions and use partitions of unity.

5-15. Let M be an $(n - 1)$ -dimensional manifold in \mathbf{R}^n . Let $M(\epsilon)$ be the set of end points of normal vectors (in both directions) of length ϵ and suppose ϵ is small enough so that $M(\epsilon)$ is also an

$(n - 1)$ -dimensional manifold. Show that $M(\epsilon)$ is orientable (even if M is not). What is $M(\epsilon)$ if M is the Möbius strip?

- 5-16. Let $g: A \rightarrow \mathbf{R}^p$ be as in Theorem 5-1. If $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable and the maximum (or minimum) of f on $g^{-1}(0)$ occurs at a , show that there are $\lambda_1, \dots, \lambda_p \in \mathbf{R}$, such that

$$(1) \quad D_j f(a) = \sum_{i=1}^n \lambda_i D_j g^i(a) \quad j = 1, \dots, n.$$

Hint: This equation can be written $df(a) = \sum_{i=1}^n \lambda_i dg^i(a)$ and is obvious if $g(x) = (x^{n-p+1}, \dots, x^n)$.

The maximum of f on $g^{-1}(0)$ is sometimes called the maximum of f subject to the **constraints** $g^i = 0$. One can attempt to find a by solving the system of equations (1). In particular, if $g: A \rightarrow \mathbf{R}$, we must solve $n + 1$ equations

$$\begin{aligned} D_j f(a) &= \lambda D_j g(a), \\ g(a) &= 0, \end{aligned}$$

in $n + 1$ unknowns a^1, \dots, a^n, λ , which is often very simple if we leave the equation $g(a) = 0$ for last. This is **Lagrange's method**, and the useful but irrelevant λ is called a **Lagrangian multiplier**. The following problem gives a nice theoretical use for Lagrangian multipliers.

- 5-17. (a) Let $T: \mathbf{R}^n \rightarrow \mathbf{R}^n$ be self-adjoint with matrix $A = (a_{ij})$, so that $a_{ij} = a_{ji}$. If $f(x) = \langle Tx, x \rangle = \sum a_{ij} x^i x^j$, show that $D_k f(x) = 2 \sum_{j=1}^n a_{kj} x^j$. By considering the maximum of $\langle Tx, x \rangle$ on S^{n-1} show that there is $x \in S^{n-1}$ and $\lambda \in \mathbf{R}$ with $Tx = \lambda x$.
 (b) If $V = \{y \in \mathbf{R}^n: \langle x, y \rangle = 0\}$, show that $T(V) \subset V$ and $T: V \rightarrow V$ is self-adjoint.
 (c) Show that T has a basis of eigenvectors.

STOKES' THEOREM ON MANIFOLDS

If ω is a p -form on a k -dimensional manifold-with-boundary M and c is a singular p -cube in M , we define

$$\int_c \omega = \int_{[0,1]^p} c^* \omega$$

precisely as before; integrals over p -chains are also defined as before. In the case $p = k$ it may happen that there is an open set $W \supset [0,1]^k$ and a coordinate system $f: W \rightarrow \mathbf{R}^n$ such that $c(x) = f(x)$ for $x \in [0,1]^k$; a k -cube in M will always be

understood to be of this type. If M is oriented, the singular k -cube c is called **orientation-preserving** if f is.

5-4 Theorem. *If $c_1, c_2: [0,1]^k \rightarrow M$ are two orientation-preserving singular k -cubes in the oriented k -dimensional manifold M and ω is a k -form on M such that $\omega = 0$ outside of $c_1([0,1]^k) \cap c_2([0,1]^k)$, then*

$$\int_{c_1} \omega = \int_{c_2} \omega.$$

Proof. We have

$$\int_{c_1} \omega = \int_{[0,1]^k} c_1^*(\omega) = \int_{[0,1]^k} (c_2^{-1} \circ c_1)^* c_2^*(\omega).$$

(Here $c_2^{-1} \circ c_1$ is defined only on a subset of $[0,1]^k$ and the second equality depends on the fact that $\omega = 0$ outside of $c_1([0,1]^k) \cap c_2([0,1]^k)$.) It therefore suffices to show that

$$\int_{[0,1]^k} (c_2^{-1} \circ c_1)^* c_2^*(\omega) = \int_{[0,1]^k} c_2^*(\omega) = \int_{c_2} \omega.$$

If $c_2^*(\omega) = f dx^1 \wedge \cdots \wedge dx^k$ and $c_2^{-1} \circ c_1$ is denoted by g , then by Theorem 4-9 we have

$$\begin{aligned} (c_2^{-1} \circ c_1)^* c_2^*(\omega) &= g^*(f dx^1 \wedge \cdots \wedge dx^k) \\ &= (f \circ g) \cdot \det g' \cdot dx^1 \wedge \cdots \wedge dx^k \\ &= (f \circ g) \cdot |\det g'| \cdot dx^1 \wedge \cdots \wedge dx^k, \end{aligned}$$

since $\det g' = \det(c_2^{-1} \circ c_1)' > 0$. The result now follows from Theorem 3-13. ■

The last equation in this proof should help explain why we have had to be so careful about orientations.

Let ω be a k -form on an oriented k -dimensional manifold M . If there is an orientation-preserving singular k -cube c in M such that $\omega = 0$ outside of $c([0,1]^k)$, we define

$$\int_M \omega = \int_c \omega.$$

Theorem 5-4 shows $\int_M \omega$ does not depend on the choice of c .

Suppose now that ω is an arbitrary k -form on M . There is an open cover \mathcal{O} of M such that for each $U \in \mathcal{O}$ there is an orientation-preserving singular k -cube c with $U \subset c([0,1]^k)$. Let Φ be a partition of unity for M subordinate to this cover. We define

$$\int_M \omega = \sum_{\varphi \in \Phi} \int_M \varphi \cdot \omega$$

provided the sum converges as described in the discussion preceding Theorem 3-12 (this is certainly true if M is compact). An argument similar to that in Theorem 3-12 shows that $\int_M \omega$ does not depend on the cover \mathcal{O} or on Φ .

All our definitions could have been given for a k -dimensional manifold-with-boundary M with orientation μ . Let ∂M have the induced orientation $\partial\mu$. Let c be an orientation-preserving k -cube in M such that $c_{(k,0)}$ lies in ∂M and is the only face which has any interior points in ∂M . As the remarks after the definition of $\partial\mu$ show, $c_{(k,0)}$ is orientation-preserving if k is even, but not if k is odd. Thus, if ω is a $(k-1)$ -form on M which is 0 outside of $c([0,1]^k)$, we have

$$\int_{c_{(k,0)}} \omega = (-1)^k \int_{\partial M} \omega.$$

On the other hand, $c_{(k,0)}$ appears with coefficient $(-1)^k$ in ∂c . Therefore

$$\int_{\partial c} \omega = \int_{(-1)^k c_{(k,0)}} \omega = (-1)^k \int_{c_{(k,0)}} \omega = \int_{\partial M} \omega.$$

Our choice of $\partial\mu$ was made to eliminate any minus signs in this equation, and in the following theorem.

5-5 Theorem (Stokes' Theorem). *If M is a compact oriented k -dimensional manifold-with-boundary and ω is a $(k-1)$ -form on M , then*

$$\int_M d\omega = \int_{\partial M} \omega.$$

(Here ∂M is given the induced orientation.)

Proof. Suppose first that there is an orientation-preserving singular k -cube in $M - \partial M$ such that $\omega = 0$ outside of

$c([0,1]^k)$. By Theorem 4-13 and the definition of $d\omega$ we have

$$\int_c d\omega = \int_{[0,1]^k} c^*(d\omega) = \int_{[0,1]^k} d(c^*\omega) = \int_{\partial I^k} c^*\omega = \int_{\partial c} \omega.$$

Then

$$\int_M d\omega = \int_c d\omega = \int_{\partial c} \omega = 0,$$

since $\omega = 0$ on ∂c . On the other hand, $\int_{\partial M} \omega = 0$ since $\omega = 0$ on ∂M .

Suppose next that there is an orientation-preserving singular k -cube in M such that $c_{(k,0)}$ is the only face in ∂M , and $\omega = 0$ outside of $c([0,1]^k)$. Then

$$\int_M d\omega = \int_c d\omega = \int_{\partial c} \omega = \int_{\partial M} \omega.$$

Now consider the general case. There is an open cover \mathcal{O} of M and a partition of unity Φ for M subordinate to \mathcal{O} such that for each $\varphi \in \Phi$ the form $\varphi \cdot \omega$ is of one of the two sorts already considered. We have

$$0 = d(1) = d\left(\sum_{\varphi \in \Phi} \varphi\right) = \sum_{\varphi \in \Phi} d\varphi,$$

so that

$$\sum_{\varphi \in \Phi} d\varphi \wedge \omega = 0.$$

Since M is compact, this is a finite sum and we have

$$\sum_{\varphi \in \Phi} \int_M d\varphi \wedge \omega = 0.$$

Therefore

$$\begin{aligned} \int_M d\omega &= \sum_{\varphi \in \Phi} \int_M \varphi \cdot d\omega = \sum_{\varphi \in \Phi} \int_M d\varphi \wedge \omega + \varphi \cdot d\omega \\ &= \sum_{\varphi \in \Phi} \int_M d(\varphi \cdot \omega) = \sum_{\varphi \in \Phi} \int_{\partial M} \varphi \cdot \omega \\ &= \int_{\partial M} \omega. \quad \blacksquare \end{aligned}$$

Problems. 5-18. If M is an n -dimensional manifold (or manifold-with-boundary) in \mathbf{R}^n , with the usual orientation, show that

$\int_M f dx^1 \wedge \cdots \wedge dx^n$, as defined in this section, is the same as $\int_M f$, as defined in Chapter 3.

5-19. (a) Show that Theorem 5-5 is false if M is not compact. *Hint:* If M is a manifold-with-boundary for which 5-5 holds, then $M - \partial M$ is also a manifold-with-boundary (with empty boundary).

(b) Show that Theorem 5-5 holds for noncompact M provided that ω vanishes outside of a compact subset of M .

5-20. If ω is a $(k-1)$ -form on a compact k -dimensional manifold M , prove that $\int_M d\omega = 0$. Give a counterexample if M is not compact.

5-21. An **absolute k -tensor** on V is a function $\eta: V^k \rightarrow \mathbf{R}$ of the form $|\omega|$ for $\omega \in \Lambda^k(V)$. An **absolute k -form** on M is a function η such that $\eta(x)$ is an absolute k -tensor on M_x . Show that $\int_M \eta$ can be defined, even if M is not orientable.

5-22. If $M_1 \subset \mathbf{R}^n$ is an n -dimensional manifold-with-boundary and $M_2 \subset M_1 - \partial M_1$ is an n -dimensional manifold-with-boundary, and M_1, M_2 are compact, prove that

$$\int_{\partial M_1} \omega = \int_{\partial M_2} \omega,$$

where ω is an $(n-1)$ -form on M_1 , and ∂M_1 and ∂M_2 have the orientations induced by the usual orientations of M_1 and M_2 . *Hint:* Find a manifold-with-boundary M such that $\partial M = \partial M_1 \cup \partial M_2$ and such that the induced orientation on ∂M agrees with that for ∂M_1 on ∂M_1 and is the negative of that for ∂M_2 on ∂M_2 .

THE VOLUME ELEMENT

Let M be a k -dimensional manifold (or manifold-with-boundary) in \mathbf{R}^n , with an orientation μ . If $x \in M$, then μ_x and the inner product T_x we defined previously determine a volume element $\omega(x) \in \Lambda^k(M_x)$. We therefore obtain a nowhere-zero k -form ω on M , which is called the **volume element** on M (determined by μ) and denoted dV , even though it is not generally the differential of a $(k-1)$ -form. The **volume** of M is defined as $\int_M dV$, provided this integral exists, which is certainly the case if M is compact. "Volume" is usually called **length** or **surface area** for one- and two-dimensional manifolds, and dV is denoted ds (the "element of length") or dA [or dS] (the "element of [surface] area").

A concrete case of interest to us is the volume element of an

oriented surface (two-dimensional manifold) M in \mathbf{R}^3 . Let $n(x)$ be the unit outward normal at $x \in M$. If $\omega \in \Lambda^2(M_x)$ is defined by

$$\omega(v, w) = \det \begin{pmatrix} v \\ w \\ n(x) \end{pmatrix},$$

then $\omega(v, w) = 1$ if v and w are an orthonormal basis of M_x with $[v, w] = \mu_x$. Thus $dA = \omega$. On the other hand, $\omega(v, w) = \langle v \times w, n(x) \rangle$ by definition of $v \times w$. Thus we have

$$dA(v, w) = \langle v \times w, n(x) \rangle.$$

Since $v \times w$ is a multiple of $n(x)$ for $v, w \in M_x$, we conclude that

$$dA(v, w) = |v \times w|$$

if $[v, w] = \mu_x$. If we wish to compute the area of M , we must evaluate $\int_{[0,1]^2} c^*(dA)$ for orientation-preserving singular 2-cubes c . Define

$$E(a) = [D_1 c^1(a)]^2 + [D_1 c^2(a)]^2 + [D_1 c^3(a)]^2,$$

$$\begin{aligned} F(a) = & D_1 c^1(a) \cdot D_2 c^1(a) \\ & + D_1 c^2(a) \cdot D_2 c^2(a) \\ & + D_1 c^3(a) \cdot D_2 c^3(a), \end{aligned}$$

$$G(a) = [D_2 c^1(a)]^2 + [D_2 c^2(a)]^2 + [D_2 c^3(a)]^2.$$

Then

$$\begin{aligned} c^*(dA)((e_1)_a, (e_2)_a) &= dA(c_*((e_1)_a), c_*((e_2)_a)) \\ &= |(D_1 c^1(a), D_1 c^2(a), D_1 c^3(a)) \times (D_2 c^1(a), D_2 c^2(a), D_2 c^3(a))| \\ &= \sqrt{E(a)G(a) - F(a)^2} \end{aligned}$$

by Problem 4-9. Thus

$$\int_{[0,1]^2} c^*(dA) = \int_{[0,1]^2} \sqrt{EG - F^2}.$$

Calculating surface area is clearly a foolhardy enterprise; fortunately one seldom needs to know the area of a surface. Moreover, there is a simple expression for dA which suffices for theoretical considerations.

5-6 Theorem. *Let M be an oriented two-dimensional manifold (or manifold-with-boundary) in \mathbf{R}^3 and let n be the unit outward normal. Then*

$$(1) \quad dA = n^1 dy \wedge dz + n^2 dz \wedge dx + n^3 dx \wedge dy.$$

Moreover, on M we have

$$(2) \quad n^1 dA = dy \wedge dz.$$

$$(3) \quad n^2 dA = dz \wedge dx.$$

$$(4) \quad n^3 dA = dx \wedge dy.$$

Proof.

Equation (1) is equivalent to the equation

$$dA(v, w) = \det \begin{pmatrix} v \\ w \\ n(x) \end{pmatrix}.$$

This is seen by expanding the determinant by minors along the bottom row. To prove the other equations, let $z \in \mathbf{R}^3_x$. Since $v \times w = \alpha n(x)$ for some $\alpha \in \mathbf{R}$, we have

$$\langle z, n(x) \rangle \cdot \langle v \times w, n(x) \rangle = \langle z, n(x) \rangle \alpha = \langle z, \alpha n(x) \rangle = \langle z, v \times w \rangle.$$

Choosing $z = e_1, e_2$, and e_3 we obtain (2), (3), and (4). ■

A word of caution: if $\omega \in \Lambda^2(\mathbf{R}^3_a)$ is defined by

$$\begin{aligned} \omega &= n^1(a) \cdot dy(a) \wedge dz(a) \\ &\quad + n^2(a) \cdot dz(a) \wedge dx(a) \\ &\quad + n^3(a) \cdot dx(a) \wedge dy(a), \end{aligned}$$

it is *not* true, for example, that

$$n^1(a) \cdot \omega = dy(a) \wedge dz(a).$$

The two sides give the same result only when applied to $v, w \in M_a$.

A few remarks should be made to justify the definition of length and surface area we have given. If $c: [0, 1] \rightarrow \mathbf{R}^n$ is differentiable and $c([0, 1])$ is a one-dimensional manifold-with-boundary, it can be shown, but the proof is messy, that the length of $c([0, 1])$ is indeed the least upper bound of the lengths

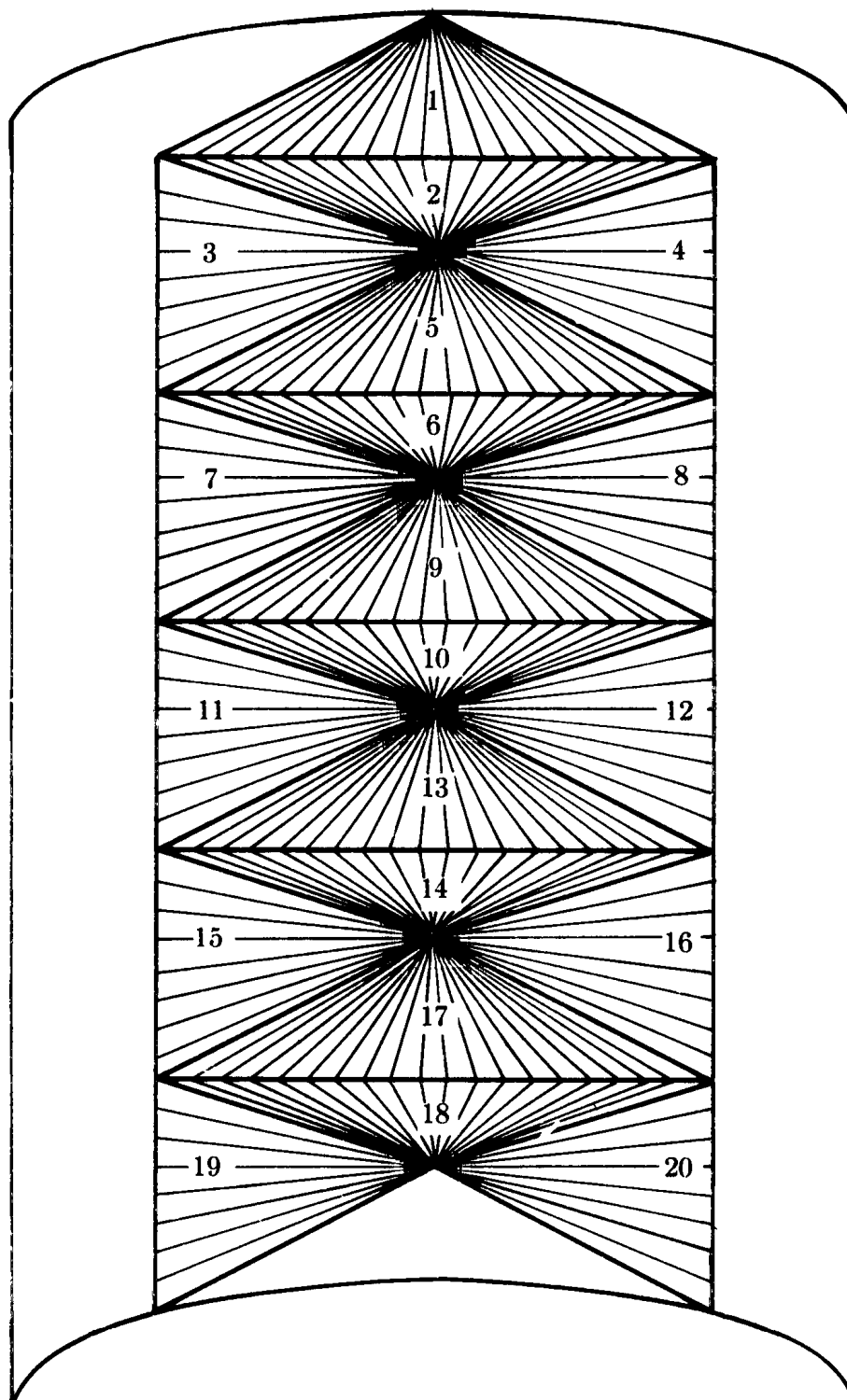


FIGURE 5-9. A surface containing 20 triangles inscribed in a portion of a cylinder. If the number of triangles is increased sufficiently, by making the bases of triangles 3, 4, 7, 8, etc., sufficiently small, the total area of the inscribed surface can be made as large as desired.

of inscribed broken lines. If $c: [0,1]^2 \rightarrow \mathbf{R}^n$, one naturally hopes that the area of $c([0,1]^2)$ will be the least upper bound of the areas of surfaces made up of triangles whose vertices lie in $c([0,1]^2)$. Amazingly enough, such a least upper bound is usually nonexistent—one can find inscribed polygonal surfaces arbitrarily close to $c([0,1]^2)$ with arbitrarily large area! This is indicated for a cylinder in Figure 5-9. Many definitions of surface area have been proposed, disagreeing with each other, but all agreeing with our definition for differentiable surfaces. For a discussion of these difficult questions the reader is referred to References [3] or [15].

Problems. 5-23. If M is an oriented one-dimensional manifold in \mathbf{R}^n and $c: [0,1] \rightarrow M$ is orientation-preserving, show that

$$\int_{[0,1]} c^*(ds) = \int_{[0,1]} \sqrt{[(c^1)']^2 + \cdots + [(c^n)']^2}.$$

- 5-24.** If M is an n -dimensional manifold in \mathbf{R}^n , with the usual orientation, show that $dV = dx^1 \wedge \cdots \wedge dx^n$, so that the volume of M , as defined in this section, is the volume as defined in Chapter 3. (Note that this depends on the numerical factor in the definition of $\omega \wedge \eta$.)
- 5-25.** Generalize Theorem 5-6 to the case of an oriented $(n - 1)$ -dimensional manifold in \mathbf{R}^n .
- 5-26.** (a) If $f: [a,b] \rightarrow \mathbf{R}$ is non-negative and the graph of f in the xy -plane is revolved around the x -axis in \mathbf{R}^3 to yield a surface M , show that the area of M is

$$\int_a^b 2\pi f \sqrt{1 + (f')^2}.$$

(b) Compute the area of S^2 .

- 5-27.** If $T: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is a norm preserving linear transformation and M is a k -dimensional manifold in \mathbf{R}^n , show that M has the same volume as $T(M)$.
- 5-28.** (a) If M is a k -dimensional manifold, show that an absolute k -tensor $|dV|$ can be defined, even if M is not orientable, so that the volume of M can be defined as $\int_M |dV|$.
- (b) If $c: [0,2\pi] \times (-1,1) \rightarrow \mathbf{R}^3$ is defined by $c(u,v) =$
- $$(2 \cos u + v \sin(u/2) \cos u, 2 \sin u + v \sin(u/2) \sin u, v \cos u/2),$$
- show that $c([0,2\pi] \times (-1,1))$ is a Möbius strip and find its area.

- 5-29. If there is a nowhere-zero k -form on a k -dimensional manifold M , show that M is orientable.
- 5-30. (a) If $f: [0,1] \rightarrow \mathbf{R}$ is differentiable and $c: [0,1] \rightarrow \mathbf{R}^2$ is defined by $c(x) = (x, f(x))$, show that $c([0,1])$ has length $\int_0^1 \sqrt{1 + (f')^2}$.
 (b) Show that this length is the least upper bound of lengths of inscribed broken lines. *Hint:* If $0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1$, then

$$\begin{aligned} |c(t_i) - c(t_{i-1})| &= \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2} \\ &= \sqrt{(t_i - t_{i-1})^2 + f'(s_i)^2 (t_i - t_{i-1})^2} \end{aligned}$$

for some $s_i \in [t_{i-1}, t_i]$.

- 5-31. Consider the 2-form ω defined on $\mathbf{R}^3 - 0$ by

$$\omega = \frac{x \, dy \wedge dz + y \, dz \wedge dx + z \, dx \wedge dy}{(x^2 + y^2 + z^2)^{\frac{3}{2}}}.$$

- (a) Show that ω is closed.
 (b) Show that

$$\omega(p)(v_p, w_p) = \frac{\langle v \times w, p \rangle}{|p|^3}.$$

For $r > 0$ let $S^2(r) = \{x \in \mathbf{R}^3: |x| = r\}$. Show that ω restricted to the tangent space of $S^2(r)$ is $1/r^2$ times the volume element, and that $\int_{S^2(r)} \omega = 4\pi$. Conclude that ω is not exact. Nevertheless we denote ω by $d\theta$ since, as we shall see, $d\theta$ is the analogue of the 1-form $d\theta$ on $\mathbf{R}^2 - 0$.

(c) If v_p is a tangent vector such that $v = \lambda p$ for some $\lambda \in \mathbf{R}$ show that $d\theta(p)(v_p, w_p) = 0$ for all w_p . If a two-dimensional manifold M in \mathbf{R}^3 is part of a generalized cone, that is, M is the union of segments of rays through the origin, show that $\int_M d\theta = 0$.

(d) Let $M \subset \mathbf{R}^3 - 0$ be a compact two-dimensional manifold-with-boundary such that every ray through 0 intersects M at most once (Figure 5-10). The union of those rays through 0 which intersect M , is a solid cone $C(M)$. The **solid angle** subtended by M is defined as the area of $C(M) \cap S^2$, or equivalently as $1/r^2$ times the area of $C(M) \cap S^2(r)$ for $r > 0$. Prove that the solid angle subtended by M is $|\int_M d\theta|$. *Hint:* Choose r small enough so that there is a three-dimensional manifold-with-boundary N (as in Figure 5-10) such that ∂N is the union of M and $C(M) \cap S^2(r)$, and a part of a generalized cone. (Actually, N will be a manifold-with-corners; see the remarks at the end of the next section.)

(b) If $r(u,v) = |f(u) - g(v)|$ show that

$$l(f,g) = \frac{-1}{4\pi} \int_0^1 \int_0^1 \frac{1}{[r(u,v)]^3} \cdot A(u,v) \, du \, dv$$

where

$$A(u,v) = \det \begin{pmatrix} (f^1)'(u) & (f^2)'(u) & (f^3)'(u) \\ (g^1)'(v) & (g^2)'(v) & (g^3)'(v) \\ f^1(u) - g^1(v) & f^2(u) - g^2(v) & f^3(u) - g^3(v) \end{pmatrix}.$$

(c) Show that $l(f,g) = 0$ if f and g both lie in the xy -plane. The curves of Figure 4-5 (b) are given by $f(u) = (\cos u, \sin u, 0)$ and $g(v) = (1 + \cos v, 0, \sin v)$. You may easily convince yourself that calculating $l(f,g)$ by the above integral is hopeless in this case. The following problem shows how to find $l(f,g)$ without explicit calculations.

5-33. (a) If $(a,b,c) \in \mathbf{R}^3$ define

$$d\Theta_{(a,b,c)} = \frac{(x-a)dy \wedge dz + (y-b)dz \wedge dx + (z-c)dx \wedge dy}{[(x-a)^2 + (y-b)^2 + (z-c)^2]^{\frac{3}{2}}}.$$

If M is a compact two-dimensional manifold-with-boundary in \mathbf{R}^3 and $(a,b,c) \notin M$ define

$$\Omega(a,b,c) = \int_M d\Theta_{(a,b,c)}.$$

Let (a,b,c) be a point on the same side of M as the outward normal and (a',b',c') a point on the opposite side. Show that by choosing (a,b,c) sufficiently close to (a',b',c') we can make $\Omega(a,b,c) - \Omega(a',b',c')$ as close to -4π as desired. *Hint:* First show that if $M = \partial N$ then $\Omega(a,b,c) = -4\pi$ for $(a,b,c) \in N - M$ and $\Omega(a,b,c) = 0$ for $(a,b,c) \notin N$.

(b) Suppose $f([0,1]) = \partial M$ for some compact oriented two-dimensional manifold-with-boundary M . (If f does not intersect itself such an M always exists, even if f is knotted, see [6], page 138.) Suppose that whenever g intersects M at x the tangent vector v of g is not in M_x . Let n^+ be the number of intersections where v points in the same direction as the outward normal and n^- the number of other intersections. If $n = n^+ - n^-$ show that

$$n = \frac{-1}{4\pi} \int_g d\Omega.$$

(c) Prove that

$$\begin{aligned} D_1\Omega(a,b,c) &= \int_f \frac{(y-b)dz - (z-c)dy}{r^3} \\ D_2\Omega(a,b,c) &= \int_f \frac{(z-c)dx - (x-a)dz}{r^3} \\ D_3\Omega(a,b,c) &= \int_f \frac{(x-a)dy - (y-b)dx}{r^3}, \end{aligned}$$

where $r(x,y,z) = |(x,y,z)|$.

(d) Show that the integer n of (b) equals the integral of Problem 5-32(b), and use this result to show that $l(f,g) = 1$ if f and g are the curves of Figure 4-6 (b), while $l(f,g) = 0$ if f and g are the curves of Figure 4-6 (c). (These results were known to Gauss [7]. The proofs outlined here are from [4] pp. 409–411; see also [13], Volume 2, pp. 41–43.)

THE CLASSICAL THEOREMS

We have now prepared all the machinery necessary to state and prove the classical “Stokes’ type” of theorems. We will indulge in a little bit of self-explanatory classical notation.

5-7 Theorem (Green’s Theorem). *Let $M \subset \mathbf{R}^2$ be a compact two-dimensional manifold-with-boundary. Suppose that $\alpha, \beta: M \rightarrow \mathbf{R}$ are differentiable. Then*

$$\begin{aligned} \int_{\partial M} \alpha dx + \beta dy &= \int_M (D_1\beta - D_2\alpha) dx \wedge dy \\ &= \iint_M \left(\frac{\partial \beta}{\partial x} - \frac{\partial \alpha}{\partial y} \right) dx dy. \end{aligned}$$

(Here M is given the usual orientation, and ∂M the induced orientation, also known as the counterclockwise orientation.)

Proof. This is a very special case of Theorem 5-5, since $d(\alpha dx + \beta dy) = (D_1\beta - D_2\alpha) dx \wedge dy$. ■

5-8 Theorem (Divergence Theorem). Let $M \subset \mathbf{R}^3$ be a compact three-dimensional manifold-with-boundary and n the unit outward normal on ∂M . Let F be a differentiable vector field on M . Then

$$\int_M \operatorname{div} F \, dV = \int_{\partial M} \langle F, n \rangle \, dA.$$

This equation is also written in terms of three differentiable functions $\alpha, \beta, \gamma: M \rightarrow \mathbf{R}$:

$$\int_M \left(\frac{\partial \alpha}{\partial x} + \frac{\partial \beta}{\partial y} + \frac{\partial \gamma}{\partial z} \right) dV = \int_{\partial M} (n^1 \alpha + n^2 \beta + n^3 \gamma) \, dS.$$

Proof. Define ω on M by $\omega = F^1 dy \wedge dz + F^2 dz \wedge dx + F^3 dx \wedge dy$. Then $d\omega = \operatorname{div} F \, dV$. According to Theorem 5-6, on ∂M we have

$$\begin{aligned} n^1 dA &= dy \wedge dz, \\ n^2 dA &= dz \wedge dx, \\ n^3 dA &= dx \wedge dy. \end{aligned}$$

Therefore on ∂M we have

$$\begin{aligned} \langle F, n \rangle dA &= F^1 n^1 dA + F^2 n^2 dA + F^3 n^3 dA \\ &= F^1 dy \wedge dz + F^2 dz \wedge dx + F^3 dx \wedge dy \\ &= \omega. \end{aligned}$$

Thus, by Theorem 5-5 we have

$$\int_M \operatorname{div} F \, dV = \int_M d\omega = \int_{\partial M} \omega = \int_{\partial M} \langle F, n \rangle \, dA. \quad \blacksquare$$

5-9 Theorem (Stokes' Theorem). Let $M \subset \mathbf{R}^3$ be a compact oriented two-dimensional manifold-with-boundary and n the unit outward normal on M determined by the orientation of M . Let ∂M have the induced orientation. Let T be the vector field on ∂M with $ds(T) = 1$ and let F be a differentiable vector field in an open set containing M . Then

$$\int_M \langle (\nabla \times F), n \rangle \, dA = \int_{\partial M} \langle F, T \rangle \, ds.$$

This equation is sometimes written

$$\int_{\partial M} \alpha dx + \beta dy + \gamma dz = \iint_M \left[n^1 \left(\frac{\partial \gamma}{\partial y} - \frac{\partial \beta}{\partial z} \right) + n^2 \left(\frac{\partial \alpha}{\partial z} - \frac{\partial \gamma}{\partial x} \right) + n^3 \left(\frac{\partial \beta}{\partial x} - \frac{\partial \alpha}{\partial y} \right) \right] dS.$$

Proof. Define ω on M by $\omega = F^1 dx + F^2 dy + F^3 dz$. Since $\nabla \times F$ has components $D_2 F^3 - D_3 F^2$, $D_3 F^1 - D_1 F^3$, $D_1 F^2 - D_2 F^1$, it follows, as in the proof of Theorem 5-8, that on M we have

$$\begin{aligned} \langle (\nabla \times F), n \rangle dA &= (D_2 F^3 - D_3 F^2) dy \wedge dz \\ &\quad + (D_3 F^1 - D_1 F^3) dz \wedge dx \\ &\quad + (D_1 F^2 - D_2 F^1) dx \wedge dy \\ &= d\omega. \end{aligned}$$

On the other hand, since $ds(T) = 1$, on ∂M we have

$$\begin{aligned} T^1 ds &= dx, \\ T^2 ds &= dy, \\ T^3 ds &= dz. \end{aligned}$$

(These equations may be checked by applying both sides to T_x , for $x \in \partial M$, since T_x is a basis for $(\partial M)_x$.)

Therefore on ∂M we have

$$\begin{aligned} \langle F, T \rangle ds &= F^1 T^1 ds + F^2 T^2 ds + F^3 T^3 ds \\ &= F^1 dx + F^2 dy + F^3 dz \\ &= \omega. \end{aligned}$$

Thus, by Theorem 5-5, we have

$$\int_M \langle (\nabla \times F), n \rangle dA = \int_M d\omega = \int_{\partial M} \omega = \int_{\partial M} \langle F, T \rangle ds. \quad \blacksquare$$

Theorems 5-8 and 5-9 are the basis for the names $\operatorname{div} F$ and $\operatorname{curl} F$. If $F(x)$ is the velocity vector of a fluid at x (at some time) then $\int_{\partial M} \langle F, n \rangle dA$ is the amount of fluid “diverging” from M . Consequently the condition $\operatorname{div} F = 0$ expresses

the fact that the fluid is incompressible. If M is a disc, then $\int_{\partial M} \langle F, T \rangle ds$ measures the amount that the fluid curls around the center of the disc. If this is zero for all discs, then $\nabla \times F = 0$, and the fluid is called *irrotational*.

These interpretations of $\operatorname{div} F$ and $\operatorname{curl} F$ are due to Maxwell [13]. Maxwell actually worked with the negative of $\operatorname{div} F$, which he accordingly called the *convergence*. For $\nabla \times F$ Maxwell proposed "with great diffidence" the terminology *rotation of F*; this unfortunate term suggested the abbreviation $\operatorname{rot} F$ which one occasionally still sees.

The classical theorems of this section are usually stated in somewhat greater generality than they are here. For example, Green's Theorem is true for a square, and the Divergence Theorem is true for a cube. These two particular facts can be proved by approximating the square or cube by manifolds-with-boundary. A thorough generalization of the theorems of this section requires the concept of manifolds-with-corners; these are subsets of \mathbf{R}^n which are, up to diffeomorphism, locally a portion of \mathbf{R}^k which is bounded by pieces of $(k - 1)$ -planes. The ambitious reader will find it a challenging exercise to define manifolds-with-corners rigorously and to investigate how the results of this entire chapter may be generalized.

- Problems. 5-34.** Generalize the divergence theorem to the case of an n -manifold with boundary in \mathbf{R}^n .
- 5-35.** Applying the generalized divergence theorem to the set $M = \{x \in \mathbf{R}^n: |x| \leq a\}$ and $F(x) = x_x$, find the volume of $S^{n-1} = \{x \in \mathbf{R}^n: |x| = 1\}$ in terms of the n -dimensional volume of $B_n = \{x \in \mathbf{R}^n: |x| \leq 1\}$. (This volume is $\pi^{n/2}/(n/2)!$ if n is even and $2^{(n+1)/2}\pi^{(n-1)/2}/1 \cdot 3 \cdot 5 \cdot \dots \cdot n$ if n is odd.)
- 5-36.** Define F on \mathbf{R}^3 by $F(x) = (0, 0, cx^3)_x$ and let M be a compact three-dimensional manifold-with-boundary with $M \subset \{x: x^3 \leq 0\}$. The vector field F may be thought of as the downward pressure of a fluid of density c in $\{x: x^3 \leq 0\}$. Since a fluid exerts equal pressures in all directions, we define the *buoyant force* on M , due to the fluid, as $-\int_{\partial M} \langle F, n \rangle dA$. Prove the following theorem. *Theorem (Archimedes).* The buoyant force on M is equal to the weight of the fluid displaced by M .

Bibliography

1. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1953.
2. Auslander and MacKenzie, *Introduction to Differentiable Manifolds*, McGraw-Hill, New York, 1963.
3. Cesari, *Surface Area*, Princeton University Press, Princeton, New Jersey, 1956.
4. Courant, *Differential and Integral Calculus*, Volume II, Interscience, New York, 1937.
5. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
6. Fort, *Topology of 3-Manifolds*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
7. Gauss, *Zur mathematischen Theorie der electrodynamischen Wirkungen*, [4] (Nachlass) Werke V, 605.
8. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.
9. Hilton and Wylie, *Homology Theory*, Cambridge University Press, New York, 1960.
10. Hu, *Homotopy Theory*, Academic Press, New York, 1959.
11. Kelley, *General Topology*, Van Nostrand, Princeton, New Jersey, 1955.

12. Kobayashi and Nomizu, *Foundations of Differential Geometry*, Interscience, New York, 1963.
13. Maxwell, *Electricity and Magnetism*, Dover, New York, 1954.
14. Natanson, *Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1955.
15. Radó, *Length and Area*, Volume XXX, American Mathematical Society, Colloquium Publications, New York, 1948.
16. de Rham, *Variétés Différentiables*, Hermann, Paris, 1955.
17. Sternberg, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

Index

- Absolute differential form, 126
- Absolute tensor, 126
- Absolute value, 1
- Algebra, Fundamental Theorem of, 105
- Alternating tensor, 78
- Analytic function, 105
- Angle, 4
 - preserving, 4
 - solid, 131
- Approximation, 15
- Archimedes, 137
- Area, 56
 - element of, 126
 - surface, 126, 127
- Basis, usual for \mathbf{R}^n , 3
- Bilinear function, 3, 23
- Boundary
 - of a chain, 97, 98
 - of a manifold-with-boundary, 113
- Boundary, of a set, 7
- Buoyant force, 137
- Cauchy Integral Formula, 106
- Cauchy Integral Theorem, 106
- Cauchy-Riemann equations, 105
- Cavalieri's principle, 62
- Chain, 97, 100
- Chain rule, 19, 32
- Change of variable, 67–72
- Characteristic function, 55
- Closed curve, 106
- Closed differential form, 92
- Closed rectangle, 5
- Closed set, 5
- Compact, 7
- Complex numbers, 104
- Complex variables, 105
- Component function, 11, 87
- Composition, 11
- Cone, generalized, 131

- Consistent choices of orientation, 117
- Constant function, 20
- Constraints, 122
- Content, 56
- Content zero, 51
- Continuous differential form, 88
- Continuous function, 12
- Continuous vector field, 87
- Continuously differentiable, 31
- Convergence, 137
- Coordinate condition, 111
- Coordinate system, 111
 - polar, 73
- Counterclockwise orientation, 134
- Cover, 7
- Cross product, 84
- Cube
 - singular, 97
 - standard n -cube, 97
- Curl, 88, 137
- Curve, 97
 - closed, 106
 - differentiable, 96
- C^∞ , 26
- Degenerate singular cube, 105
- Derivative, 16
 - directional, 33
 - partial, 25
 - higher-order (mixed), 26
 - second-order (mixed), 26
- Diffeomorphism, 109
- Differentiable function, 15, 16, 105
 - continuously, 31
- Differentiable curve, 96
- Differentiable differential form, 88
 - on a manifold, 117
- Differentiable vector field, 87
 - on a manifold, 115
- Differentiable = C^∞ , 88
- Differential, 91
- Differential form, 88
 - absolute, 126
 - closed, 92
 - continuous, 88
 - differentiable, 88
 - exact, 92
- Differential form, on a manifold, 117
 - differentiable, 117
- Dimension
 - of a manifold, 109
 - of a manifold-with-boundary, 113
- Directional derivative, 33
- Distance, 4
- Divergence of a field, 88, 137
- Divergence Theorem, 135
- Domain, 11
- Dual space, 5
- Element of area, 126
- Element of length, 126
- Element of volume, *see* Volume element
- End point, 87
- Equal up to n th order, 18
- Euclidean space, 1
- Exact differential form, 92
- Exterior of a set, 7
- Faces of a singular cube, 98
- Field, *see* Vector field
- Form, *see* Differential form
- Fubini's Theorem, 58
- Function, 11
 - analytic, 105
 - characteristic, 55
 - component, 11, 87
 - composition of, 11
 - constant, 20
 - continuous, 12
 - continuously differentiable, 31
 - C^∞ , 26
 - differentiable, 15, 16, 105
 - homogeneous, 34
 - identity, 11
 - implicitly defined, 41
 - see also* Implicit Function Theorem
 - integrable, 48
 - inverse, 11, 34–39
 - see also* Inverse Function Theorem
 - projection, 11

- Fundamental Theorem of Algebra, 105
- Fundamental Theorem of Calculus, 100–104
- Gauss, 134
- Generalized cone, 131
- Grad f , 96
- Graph, 11, 115
- Green's Theorem, 134
- Half-space, 113
- Heine-Borel Theorem, 7
- Homogeneous function, 34
- Homotopy, 108
- Identity function, 11
- Implicit Function Theorem, 41
- Implicitly defined function, 41
- Incompressible fluid, 137
- Independence of parameterization, 104
- Induced orientation, 119
- Inequality, *see* Triangle inequality
- Inner product, 2, 77
 - preserving, 4
 - usual, 77, 87
- Integrable function, 48
- Integral, 48
 - iterated, 59, 60
 - line, 101
 - lower, 58
 - of a form on a manifold, 123–124
 - of a form over a chain, 101
 - over a set, 55
 - over an open set, 65
 - surface, 102
 - upper, 58
- Integral Formula, Cauchy, 106
- Integral Theorem, Cauchy, 106
- Interior of a set, 7
- Inverse function, 11, 34–39
- Inverse Function Theorem, 35
- Irrotational fluid, 137
- Iterated integral, 59, 60
- Jacobian matrix, 17
- Jordan-measurable, 56
- Kelvin, 74
- Lac* locus, 106
- Lagrange's method, 122
- Lagrangian multiplier, 122
- Leibnitz's Rule, 62
- Length, 56, 126
 - element of, 126
- Length = norm, 1
- Limit, 11
- Line, 1
- Line integral, 101
- Linking number, 132
- Liouville, 74
- Lower integral, 58
- Lower sum, 47
- Manifold, 109
- Manifold-with-boundary, 113
- Manifold-with-corners, 131, 137
- Mathematician (old style), 74
- Matrix, 1
 - Jacobian, 17
 - transpose of, 23, 83
- Maxima, 26–27
- Measure zero, 50
- Minima, 26–27
- Möbius strip, 119, 120, 130
- Multilinear function, 23, 75
- Multiplier, *see* Lagrangian multiplier
- Norm, 1
- Norm preserving, 4
- Normal, *see* Outward unit normal
- Notation, 3, 44, 89
- One-one (1-1) function, 11
- One-sided surface, 121
- Open cover, 7
- Open rectangle, 5
- Open set, 5
- Orientable manifold, 119
- Orientation, 82, 119
 - consistent choices of, 117
 - counterclockwise, 134
 - induced, 119
 - usual, 83, 87, 121

- Orientation-preserving, 118, 123
- Oriented manifold, 119
- Orthogonal vectors, 5
- Orthonormal basis, 77
- Oscillation, 13
- Outward unit normal, 119, 120

- Parameterization, independence of, 104
- Partial derivative, 25
 - higher-order (mixed), 26
 - second-order (mixed), 26
- Partition
 - of a closed interval, 46
 - of a closed rectangle, 46
 - of unity, 63
- Perpendicular, 5
- Plane, 1
- Poincare Lemma, 94
- Point, 1
- Polar coordinate system, 73
- Polarization identity, 5
- Positive definiteness, 3, 77
- Product, *see* Cross product, Inner product, Tensor product, Wedge product
- Projection function, 11

- Rectangle (closed or open), 5
- Refine a partition, 47
- Rotation of F , 137

- Sard's Theorem, 72
- Self-adjoint, 85
- Sign of a permutation, 78
- Singular n -cube, 97
- Solid angle, 131
- Space, 1
 - see also* Dual space, Euclidean space, Half-space, Tangent space
- Sphere, 111
- Standard n -cube, 97
- Star-shaped, 93

- Stokes' Theorem, 102, 124, 135
- Subordinate, 63
- Subrectangles of a partition, 46
- Surface, 127
- Surface area, 126, 127
- Surface integral, 102
- Symmetric, 2, 77

- Tangent space, 86, 115
- Tangent vector, 96
- Tensor, 75
 - absolute, 126
 - alternating, 78
- Tensor product, 75
- Torus, 115
- Transpose of a matrix, 23, 83
- Triangle inequality, 4

- Unit outward normal, 119, 120
- Upper integral, 58
- Upper sum, 47
- Usual, *see* Basis, Inner product, Orientation

- Variable
 - change of, 67–72
 - complex, *see* Complex variables
 - function of n , 11
 - independent of the first, 18
 - independent of the second, 17
- Vector, 1
 - tangent, 96
- Vector field, 87
 - continuous, 87
 - differentiable, 87
 - on a manifold, 115
 - continuous, 87
 - differentiable, 115
- Vector-valued function, 11
- Volume, 47, 56, 126
- Volume element, 83, 126

- Wedge product, 79
- Winding number, 104

Addenda

1. It should be remarked after Theorem 2-11 (the Inverse Function Theorem) that the formula for f^{-1} allows us to conclude that f^{-1} is actually continuously differentiable (and that it is C^∞ if f is). Indeed, it suffices to note that the entries of the inverse of a matrix A are C^∞ functions of the entries of A . This follows from "Cramer's Rule": $(A^{-1})_{ji} = (\det A^{ij})/(\det A)$, where A^{ij} is the matrix obtained from A by deleting row i and column j .

2. The proof of the first part of Theorem 3-8 can be simplified considerably, rendering Lemma 3-7 unnecessary. It suffices to cover B by the interiors of closed rectangles U_i with $\sum_{i=1}^\infty v(U_i) < \epsilon$, and to choose for each $x \in A - B$ a closed rectangle V_x , containing x in its interior, with $M_{V_x}(f) - m_{V_x}(f) < \epsilon$. If every subrectangle of a partition P is contained in one of some finite collection of U_i 's and V_x 's which cover A , and $|f(x)| \leq M$ for all x in A , then $U(f, P) - L(f, P) < \epsilon v(A) + 2M\epsilon$.

The proof of the converse part contains an error, since $M_s(f) - m_s(f) \geq 1/n$ is guaranteed only if the interior of S intersects $B_{1/n}$. To compensate for this it suffices to cover the boundaries of all subrectangles of P with a finite collection of rectangles with total volume $< \epsilon$. These, together with \mathcal{S} , cover $B_{1/n}$, and have total volume $< 2\epsilon$.

3. The argument in the first part of Theorem 3-14 (Sard's Theorem) requires a little amplification. If $U \subset A$ is a closed rectangle with sides of length l , then, because U is compact, there is an integer N with the following property: if U is divided into N^n rectangles, with sides of length l/N , then $|D_j g^i(w) - D_j g^i(z)| < \varepsilon/n^2$ whenever w and z are both in one such rectangle S . Given $x \in S$, let $f(z) = Dg(x)(z) - g(z)$. Then, if $z \in S$,

$$|D_j f^i(z)| = |D_j g^i(x) - D_j g^i(z)| < \varepsilon/n^2.$$

So by Lemma 2-10, if $x, y \in S$, then

$$\begin{aligned} |Dg(x)(y - x) - g(y) + g(x)| &= |f(y) - f(x)| < \varepsilon|x - y| \\ &\leq \varepsilon \sqrt{n} (l/N). \end{aligned}$$

4. Finally, the notation $\Lambda^k(V)$ appearing in this book is incorrect, since it conflicts with the standard definition of $\Lambda^k(V)$ (as a certain quotient of the tensor algebra of V). For the vector space in question (which is naturally isomorphic to $\Lambda^k(V^*)$ for finite dimensional vector spaces V) the notation $\Omega^k(V)$ is probably on the way to becoming standard. This substitution should be made on pages 78-85, 88-89, 116, and 126-128.